AI and extremism in social networks

Exploring the role of non-human actors in counterinsurgencies against radicalization

TEDLA YENEAKAL

Master Thesis, Humanities, Digital Culture.

UNIVERSITY OF BERGEN

(20.11.2019)

Key words: [violent extremism, social media, radicalization, AI, cognitive assemblages, social movements, chatbots, moral agents]

Candidate no. 101

Sammendrag

Studien utforsker hvordan midler som kunstig intelligens, AI- drevne chatbots, kan være kilder man kan regne med som moralske aktører på digitale plattformer og som kan være identifiserbare opprørsmodeller til bekjempelse av ekstremistiske og voldsforherligende ytringer på sosiale medieplattformer. Fremveksten av digital nettverkskommunikasjon har lettet prosessen med sosiale bevegelser, noe fenomenet «Den arabiske våren» tydelig demonstrerer. Sosiale medier har vært et verdifullt verktøy når det gjelder å utvikle kollektive identiteter med en felles ideologi for å fremme et bestemt mål eller en sak og gi alternative plattformer for undertrykte samfunn. Imidlertid forblir virkningen og konsekvensene av sosiale medier i samfunn der maktbalansen forrykkes gjennom fundamentale endringer et bekymringsfullt fenomen. Radikaliserte individer og grupper har også hevdet sin tilstedeværelse på sosiale medieplattformer gjennom å fremme fordommer, hat og vold. Ekstremistiske grupper bruker ulike taktikker for å utøve makten sin på disse plattformene. Bekjempelsen av voldelig ekstremisme på sosiale medieplattformer blir som regel ikke koordinert av aktuelle aktører som regjeringer, sosiale medieselskaper, FN eller andre private organisasjoner. I tillegg har fremdeles ikke forsøk på å konstituere AI til bekjempelse av voldelig ekstremisme blitt gjennomført, men lovende resultater har blitt oppnådd gjennom noen initiativer. Prosjektet som en 'case study' ser på den nylige reformen i Etiopia som ble gjennomført av Nobels fredsprisvinner 2019 Abiy Ahmed etter at han tiltrådte som statsminister i Etiopia i april 2018. Etter flere tiår med undertrykkelse har den nye maktovertakelsen der det politiske rommet ble åpnet opp og ytringsfrihet ble tillatt, uventet ført til et skred av etniske gruppers polarisering. Nye etno-ekstremister har dukket frem fra alle kriker og kroker av landet og også fra sin tilværelse i diaspora. Studien ser videre på hvilken rolle sosiale medier til tider spiller ved direkte å presse på for å påvirke til og dermed forårsake voldelige handlinger på grasrota.Ved å bruke en kvalitativ forskningsmetode for ustrukturerte intervjuer med etiopiske brukere av sosiale medier, journalister og aktivister, identifiserer studien kjerneaspektene ved konfliktene og foreslår initiativer som kan brukes til å motvirke voldelig etnisk ekstremisme. Ved å bruke relevant litteratur ser prosjektet videre på innarbeidelsen av kunstig intelligens (AI) i «moralske handlinger» på sosiale medier og hvordan den kan utformes slik at den av seg selv kan ta i bruk moralske beslutningsevner i nettverket. I tillegg ser studien på mulighetene videre for bekjempelse av voldelig ekstremisme og skisserer den spesifikke rollen ikke menneskelige aktører som profesjonelle troll og bots på sosiale medier bør spille for å slåss mot radikalisering som kan føre til voldelige handlinger.

Abstract

The study explores how artificial agents such as AI-powered chatbots can be fully accountable sources of moral action in digital platforms and be used as identifiable counter insurgency models against violent extremism on social media platforms. The emergence of digital networked communications have facilitated the process of social movements, as evident in examples such as the Arab Spring. Social media has been a valuable tool in terms of developing collective identities with a common ideology to promote a specific agenda or cause and in providing an alternative communication platform for repressed societies. However, the impact and consequences of social media in fundamentally changing power relations in society remains a concern. Radicalized individuals and groups have also used these platforms to promote bigotry, hate and violence. Extremists use several tactics to yield their power in these platforms. Counter-insurgency efforts are often not coordinated among relevant actors like governments, social media companies, the UN and other private organizations. In addition, efforts to fully constitute AI in counter-insurgency against violent extremism have not yet occurred, but promising results are being obtained from some initiatives. The project as a case study looks into the recent reform in Ethiopia being carried out by the 2019 Noble Peace Prize winner Abiy Ahmed since he took office in April 2018. After decades of repression, opening the political space and freedom of the press in Ethiopia has unexpectedy led to a surge of conflicts between polarized groups on ethnic lines. It has created ethno-extremists from all corners at times directly pushing for and impacting violent actions on the ground on social media. Using a qualitative research method of unstructured interviews with Ethiopian social media users, journalists and activists, the study identifies the core aspects of the conflicts and suggests initiatives that could be used to counter violent ethnic extremism. Further, using relevant literature, the project looks into the incorporation of Artificial Intelligence (AI) in 'moral actions' on social media and how they can be designed to inherently adopt moral decision abilities in the network. The study in addition recommennds a way forward for counterinsurgency efforts against violent extremism and outlines the specific role non-human actors such as professional trolls and bots on social media can play to battle radicalization that may lead to violence.

Contents

Chapter One

1. Introduction

The project explores how artificial agents can be fully accountable sources of moral action in digital platforms and how these agents can be able to function as ideological agents in their own terms; serving to solve complicated social problems like challenging narratives of violent extremist groups and individuals on social media, manage to disrupt these extremist groups' strategies of recruiting young members to join, acting as compromisers and moral mediators online etc.

Counter online strategies against radicalization on social media platforms have been deployed by governments and international organizations in the last decade. For the obvious reason of security, these strategies are often hidden and are not transparent. In the absence of this transparency, it is often difficult for researchers to study the productivity of these counter strategies. On the other hand, how extremists use social media as means of communication needs a broad study on its own as well as it's a daunting task to find out if extremists are using other parallel platforms, although possible.

Since this project involves the study of artificial agents, it requires looking into Artificial Intelligence (AI) that "is carried out on a task-by-task basis, with dialogues systems- such as the chatterbot, which carries out a conversation with a human over teletype- representing one of many possible tasks," (Ryan, Emerson, and Robertson 2014, 23)

This is particularly of interest to this project as it focuses on the communication aspects of the future of Artificial Intelligence and the cultural impacts of changes in the development of communication technologies considering the development of chatbots that can be programmed to be capable of acting as moral agents independently, specifically able to challenge extremist narratives that may lead to violence.

Can AI, social bots be developed to manage, negotiate, compromise, and perhaps act as online mediators to solve issues beyond borders regardless of cultural, language, religious or political differences? These are interesting questions to ponder on but need theoretical backing, as they sketch a bit farfetched notion of reality in their claim that AI can serve as social experts within their own right.

The project explores the following research question:

Can non-human actors such as artificial moral agents be effective as identifiable counter insurgency models in battling social media radicalization that may lead to violence?

1.1 Definitions of terms and relevant theoretical frameworks

1.1.1 Radicalization

Radicalization is a contested concept in academia. Before we can explore the roles and impacts of digital moral agents in fighting online radicalization on social media platforms, there is a need to investigate this contested theory. Rik Coolsaet, a Belgian expert who was part of an expert group on violent radicalization established by the European Commission to study the problem, describes the very notion of radicalization as 'ill-defined, complex and controversial." (Coolsaet 2015)

Scholars agree at large that the lack of clarity and consensus regarding many key concepts such as terrorism, radicalization, extremism, etc... – Ill-defined and yet taken for granted – still present an obstacle that needs to be overcome.

However, "nobody thinks or acts in a vacuum", and whilst "radicalization" is a lacking label it is one that currently guides policy and research on radical and violent actions, argues editor of *RadicalisationResearch.org* Matthew Francis in an article entitled "What Causes Radicalization?" (Francis 2012) This project utilizes radicalization as one of the theoretical frameworks bearing in mind the controversial nature of meaning of terms such as radicalization, de-radicalization and counter-radicalization.

In an in-depth literature review, entitled "Radicalization, De-Radicalization, Counter-Radicalization: A Conceptual Discussion and Literature Review", Alex P. Schmid explores the mentioned terms and the discourses surrounding them. In this research paper, Schmid discovers the relationship between radicalization, extremism and terrorism. He notes for example, historically, 'radicalism' – contrary to 'extremism' – does not necessarily have negative connotations, nor is it a synonym for terrorism. (Schmid 2013)

Similar to Coolsaet, Schmid acknowledges that the concept of radicalization is not as solid and clear as many seem to take for granted. He agrees with the working definition of the mentioned

European Commission expert group tasked to analyze the state of academic research on radicalization to violence that notes "radicalization is a context-bound phenomenon par excellence. Global, sociological and political drivers matter as much as ideological and psychological ones." The expert group devised a pertinent working definition of violent radicalization, "socialization to extremism which manifests itself in terrorism." (Schmid 2013)

Schmid argues that both extremism and radicalism can only be properly assessed in relation to what is mainstream political thought in each period. He further explores in detail about radicalization. It proposes to consider radicalization not only on the micro-level of "vulnerable individuals" but also on the meso-level of the "radical milieu" and the macro-level of "radicalizing public opinion and political parties". He also suggests the importance of acknowledging how the terms "radicalism" and "extremism" are alternatively used in documents in "manifesting a closed mind and distinct willingness to use violence against civilians" both standing at some distance from mainstream political thinking. (Schmid 2013, 12)

This framework is used at large for this paper, as Schmid conceptualizes radicalization as a process that can occur on both sides of conflict dyads and challenges several widespread assumptions. The final section is the most relevant for my research as it specifically examines various counter-radicalization programs. Schmid defines 'de-radicalization' as "programs that are generally directed against individuals who have become radical with the aim of reintegrating them into society or at least dissuading them from violence." (Schmid 2013, 42)

However, just like radicalization Schmid underscores the lack of conceptual clarity surrounding the discourse of de-radicalisation. As de-radicalization is often understood as any effort aimed at preventing radicalisation from happening. The other perspective of de-radicalisation focuses more on de-programming of those already radicalized.

Schmid quotes Froukje Demant and her colleagues to explain the concept of "de-programming" that states; "It is the process of becoming less radical. This process of becoming less radical applies both to behavior and beliefs. With regard to behavior, this primarily involves the cessation of violent actions. With regard to beliefs, this involves an increase in confidence in the system, a desire to once more be a part of society, and the rejection of non-democratic means. […] In general, the de-radicalization of behavior is linked with the de-radicalisation of beliefs." (Schmid 2013, 41)

Schmid further notes the importance of looking into what is incorporated in much of the literature under 'de-radicalisation'. He reminds us that the notion 'de-radicalization is the opposite of radicalization' does not seem to apply universally. Schmid notes, "this is reflected in the wide variety of measures and objectives advocated for de-radicalisation of individuals and groups, such as: amnesty; counselling; deprogramming; dialogue; demobilization; disbandment; disengagement; reconciliation; reintegration." (Schmid 2013, 192)

The aspects of radicalization and de-radicalization as well as the notion of such related terms to counter violent extremism on social media are treated critically throughout this study bearing in mind the contentious nature of these terms.

On the other hand, pressing challenges that face societies across the globe is "how to balance the precious freedom and connective power of the internet while mitigating the harms that digital technologies can pose." (Parker, Boyer, and Gatewood 2018) Digital tools are being greatly misused leading to growing threats of for instance, information manipulation, trolling, disseminating extremist content, undercut social cohesion, catalyze political polarization and undermine trust between groups and within institutions (the case study of Ethiopia in chapter four will dwell more on this). Disinformation and conspiracy theories threaten democratic systems and disrupt the ability to respond effectively to civic challenges, while online hate threatens our relations with each other.

Institute for Strategic Dialogue (ISD) is an organization founded in 2006 that works to understand and innovate technological assisted responses to the rising tide of polarization, hate and extremism of all forms. ISD combines anthropological research and know-how in international extremist movements as well as an advanced digital analysis capability that tracks hate, disinformation and extremism online. (Institute for Strategic Dialogue (ISD) 2006)

ISD has an initiative that provides an experimental, data and technology driven approach in counter-terrorism efforts. For example, ISD's initiative involved using Natural Language Processing (NLP) and Machine Learning technology, they were able to identify individuals engaging with extremist content and publicly espousing violent and dehumanizing ideas in both the Extreme Right and Islamist camps, at scale in political discourses. To confront these challenges, ISD argues that action is required from government, civil society and technology companies. Yet it is individuals who sit at the heart of these challenges.

One of the proposed frameworks for combatting the harms that digital technologies pose is the Theory of Change; that is designed to give young people the capacity they need to stay safe

online, increase their resilience to antisocial behavior, hate and extremism online, and become positive online citizens. The project involved developing the skills of participants' media literacy, critical thinking and digital citizenship skills specific to the national and international challenges of online hate, prejudice and intolerance in the countries of delivery. It also aims to increase participants understanding of propaganda, fake news, biased writing and the arguments and techniques used by content creators to manipulate people online and suggest ways of recognizing and challenging online hate speech. (Institute for Strategic Dialogue (ISD) 2006)

One aspect of intervention deployed by ISD was to develop a semi-automated identification methodology that can accurately identify individuals who are publicly expressing signs of ideologically inspired hatred and violent sentiment towards others on social media; which involved identification of Facebook accounts which were repeatedly engaging with Facebook pages associated with the extreme right or Islamist extremism, or which tended to attract individuals expressing violently extreme viewpoints.

The ISD also applied an approach that combined machine learning and a Natural Language Processing (NLP) algorithm to identify people who appeared to be using violent and dehumanizing language against other groups of people on these pages. Such approaches are considered for this project bearing in mind ISD's project focuses more on extreme right and Islamist extremism, while my case study looks into extremism from a specific local situation, mainly ethnic, political and perhaps some religious extremist disputes amongst Ethiopian social media users.

Schmid's second policy recommendation relates to the role of counter narratives to those of notable violent extremist groups such as ISIS or radical white extremist groups that have caused actual fatalities and destructions on civilians. In addition, Schmid identifies 'credibility and legitimacy' as core ingredients in counter-radicalization and counter-terrorism efforts that governments need to incorporate in their narratives, which in the eyes of domestic and foreign publics be markedly better than extremist parties and terrorist organizations.

The recommendations also include a re-conceptualization of radicalization that acknowledges certain forms of violent resistance to political oppression that are illegal under certain national laws but accepted under international humanitarian law. For example, the mobilization of Ethiopian Facebook activists to protest against a government that arbitrarily detains oppositions, critical journalists or takes harsh measures against dissents for exercising their freedom of speech can be considered illegal under the infamous and disputed Ethiopian anti-

terrorist proclamation that has been used to stifle dissents under the administration of Prime Minister Meles Zenawi. Although it would be condemned internationally as a gross violation of human rights.

Speaking of counter narrative efforts against radicalized groups, the ISD report explains the term Echo Chamber- "a metaphorical description of a situation in which beliefs are amplified or reinforced by communication and repetition inside a closed system." By visiting an 'Echo Chamber', people are able to seek out information which reinforces their existing views, potentially as an unconscious exercise of confirmation bias. (Institute for Strategic Dialogue (ISD) 2006) This may increase political and social polarization and extremism. As we will see in the examples in chapter two, in a broader spectrum of political action, it is not only non-state actors but state actors can radicalize too. Schmid highlights that in efforts related to de-radicalization, dis-engagement and counter radicalization, it is difficult to have a general framework that works for all situations but instead local contexts are varied and thus requires a specific look at individual scenarios.

However, such insurgency models are mainly dependent on human actors that have deployed offline intervention method to the social media domain, with programs providing an opportunity for individuals showing clear signs of radicalization to meet and engage with someone that can support their exit from hate. Traditional models of moral action and responsibility were developed for the kinds of actions performed by an individual that have directly visible consequences. (Noorman 2018) This is despite today's society attributions of responsibility to an individual or a certain group of individuals are intertwined with the artifacts with which they interact as well as with intentions and actions of other human agents these artifacts mediate.

Because of the increasing complexity of digital technologies and advances in AI, there is a need for a different kind of analysis of who can be responsible and what it means to be morally responsible. Philosophers like Daniel Dennett for instance suggests that a computer could be held morally responsible if it concerned a higher order "intentional computer system." According to him, an intentional system is one that can be predicted and explained by attributing beliefs and desires to it as well as rationality. Dennet in the same article entitled "Intentional System Theory" explains that primarily the purpose of the framework is to be used in analysis of meanings of such everyday terms as 'believe', 'desire', 'expect, 'decide' and 'intend,' the terms of psychology that we use to interpret, explain and predict the behavior of other human

beings, animals, some artifacts such as robots and computers as well as ourselves. (Dennett 2009) Dennett's theory is certainly relevant for this research. I will in the next two sections further review the idea of moral agency that is not restricted to human beings.

1.1.2. Moral actions in digital platforms

The idea of moral development and the sense of fair-mindedness carried within every human being has been studied for years. Researchers such as Jean Piaget since the 1930's and Lawrence Kohlberg in the 1950's and several others have done studies in cognitive psychology that constitutes the steps that are necessary in the transformation of doing the 'right' thing and models of moral behavior that entail processes necessary to ensure a moral act in social situations. For example, Narvaez. D & Rest J. R., present the four components of acting morally. Moral behavior and moral development present the four component model of moral behavior; internal processes the model deems necessary for a moral act to ensue.

These four major units of analysis are namely: moral sensitivity (involves consideration of what actions are possible, who and what might be affected by each possible action and how the involved parties react to possible outcomes.); moral judgement (involves reasoning about the possible actions and deciding which is most moral or ethical.); moral motivation (involves prioritizing what is considered to be the most moral or ethical action over others and being intent upon following that course.); and moral action (combines the strength of will the social and psychological skills necessary to carry out the intended course of action. It is dependent both on having the requisite skills and on persisting in the face of any obstacles or challenges to the action that may arise.) (Schweigert 2016, 4)

"Each process constitute cognitive, affective and behavioral aspects that function together in fostering the completion of a moral action." (Schweigert 2016, 3) Narvaez & Rest point out that these four components represent processes that together produce a moral act; "they are not character traits or virtues that make up an ideal moral person but rather they are the major units of analysis in tracking how a particular course of action came about in the particular situation." (Schweigert 2016, 4)

It is also relevant to note here in connection to moral principles, Immanuel Kant's concept of 'Categorical Imperative (CI)' where he argues that CI is 'the supreme principle of morality; that is a standard of rationality." Kant characterized the CI as an 'objective, rationally necessary

and unconditional principle that we must always follow despite any natural desires or inclinations we may have to the contrary. All specific moral requirements, according to Kant are justified by this principle, "which means that all immoral actions are irrational because they violate the CI." (Johnson and Cureton 2019)

Other philosophers such as Hobbes, Locke and Aquinas, had also argued that "moral requirements are based on standards of rationality." Although Kant agreed with his predecessors that an analysis of practical reason reveals the requirement that rational agents must conform to instrumental principles. He however argued that conformity to the CI (a non-instrumental principle), and hence to moral requirements themselves, can nevertheless be shown to be essential to rational agency. This is based on his doctrine that the fundamental principle of morality-the CI- is none other than the law of autonomous will. According to Kant, each of us have in us the presence of self-governing reason that offers decisive grounds for viewing each as possessed of equal worth and deserving of equal respect. (Johnson and Cureton 2019)

Although models like the four-component model of moral behavior aide to research moral action in digital platforms, they are far from addressing the challenges computing poses to conventional notions of moral responsibility. For example, actions related to the dissemination of hate speeches by violent extremist groups on social media and insurgency methods to combat such actions demands looking into 'moral action' and identifying in action what that really means on these digital platforms. Various authors that have studied morality in digital platforms suggest different ways in which these can be addressed; by reconsidering the idea of moral agency, by rethinking the concept of moral responsibility." (Wallach and Allen 2011, 64)

(Wallach and Allen 2011) argue that moral responsibility is about human actions and its intentions and consequences. "Generally speaking a person or a group of people is morally responsible when their voluntary actions have morally significant outcomes that would make it appropriate to blame or praise them." (Wallach and Allen 2011, 78) The person or group that performs the action and causes something is referred to as the 'agent' in this project.

In addition, digital technologies affect the choices humans have and how they make them and complicate the determination of whether someone is free to act in light of growing automation of decision making processes and control. (Fischer and Plessow 2015) argue that automation only helps to centralize and increase control over multiple processes for those in charge but limits the unrestricted power of human operators on the lower-end of the decision making chain.

B.J. Fogg in "Persuasive Technology: Using Computers to Change at We Think and Do" outlines the development of technological artifacts that can enforce and persuade humans to act in a "morally desirable behavior". He is not particularly concerned with morality per say but in general on the era of "persuasive technology of interactive computing systems designed to change people's attitudes and behaviors." (Fogg 2002, 1)

Fogg coined the term "captology" an acronym based on the phrase "computers as persuasive technologies" briefly stated, captology focuses on the design, research, and analysis of interactive computing products created for the purpose of changing people's attitudes or behaviors. (Fogg 2002, 4) Critics of the field of Persuasive Technology argue that purposely developing technology to enforce morally desirable behavior discards the democratic principles of society as it deprives humans of their ability and rights to make a decision that is carefully weighed, considered and studied and it is intentional and designed to act voluntarily. They argue that if humans are not acting freely, their actions cannot be considered moral. However, scholars such as Verbeek counter argue that technological artifacts have already set conditions for actions that humans are able to perform, indicating to the rules, regulations and a host of technological artifacts. He adds that "technological artifacts as active mediators, affect the actions and experiences of humans but they do not determine them." (Peterson and Spahn 2011)

The framework of the study of Persuasive Technology is relevant for this project as its core purpose is to study how non-human actors can be effective as insurgency models in battling social media radicalization. One aspect of interest to use Persuasive Technology is to be enabled to persuade and counter extremist rhetoric or hate speech on social media that may lead to violence. How can artificial moral agents be able to be used as mediators on their own terms and be able to perform the named tasks? In the next section, I look at the idea of using Artificial Intelligence (AI) as moral agents and the concept of cognitive assemblages.

1.1.3. Artificial moral agents and Chatbots

"Chatbots are one class of intelligent, conversational software agents activated by natural language input (which can be in the form of text, voice, or both). They provide conversational output in response, and if commanded, can sometimes also execute tasks." (Radziwill and Benton 2017)

In an academic article "Evaluating Quality of Chatbots and Intelligent Conversational Agents", (Radziwill and Benton 2017) present a literature review of quality issues and attributes that relate to the contemporary issue of chatbot development. They also elaborate on terms such as intelligent agents as well as on chatbots that can also be used to engineer social harm, for example spread rumors, misinformation, or attack people for posting their thoughts online. The theoretical framework is highly relevant to this research as it will attempt to study the role of moral agents in social media platforms that can engineer social good instead of harm.

Moreover, the mentioned article looks into how chatbot technologies have existed since the 1960s and now have advanced to the level of being able to be trained and to implement. The authors note that plentiful open source code, widely available development, and implementation option via Software as a Service (SaaS) have enabled machines to perform a complicated task such as acting as moral agents in their own terms or as programmed by external input. (Radziwill and Benton 2017)

In Social media, some social bots were developed to behave like an ordinary human user, able to post and share at reasonable hours of the day or limit the amount of information they share and other competences. According to (Curtis 2014), more than 50% of Internet traffic is generated by bots, while (Varol et al. 2017) have shown that 15% of twitter accounts are controlled by bots. Other researches for instance, (Mares and Moats 2015) have also emphasized bots capacity to shape Social Media online debates. "Bots contribute to making meaning when they are deviant or impact negatively on social order in social media." (De Paoli 2017)

De Paoli in this article introduces the concept of what he called Ordering Turing Tests: "Sort of Turning Tests proposed by social actors for purposes of deviant behavior, a method for labelling deviance where social actors can use the test to tell apart rule-abiding humans and rule-breaking bots." (De Paoli 2017) I will look more into this theory when I discuss counter-insurgency models against radicalization of use in the case study of Ethiopia.

A bot-generated internet persona is increasingly difficult to distinguish from a live human for an ordinary social media user. Burkhardt points out people who are unaware they are interacting with a bot can easily be supplied with false information and site a research from the Communications of the Association for Computing Machinery in 2016, whereby "more than 20% of authentic Facebook users accept friend requests indiscriminately making them vulnerable for bots to infiltrate a network of social media users." (Burkhardt, 2018)

This perhaps fuels other questions related to my research question, that is: if Artificial Intelligence is deployed in moral agents to battle online radicalization, how do we measure the quality of their performance? For instance, in the absence of a common consensus amongst policy makers about the use of terms such as terrorism and radicalization; how can artificial moral agents be able to cope with such dilemmas? Who is an extremist? Where do we draw the line for someone is in the process of being radicalized without violating their fundamental democratic rights? The project looks to approach such questions in a critical manner.

In an article, Luciano Floridi and J.W. Sanders, "On the Morality of artificial agents" introduce an approach that focuses directly on "mind-less morality" able to avoid many of the concerns raised by Artificial Intelligence. The vital component of their approach is what they referred to as the "Method of Abstraction" for analyzing the level of abstraction at which an agent is considered to act. The level of abstraction is determined by the way in which one chooses to describe, analyze, and discuss a system and its context. The "Method of Abstraction" is explained in terms of an "interface" or set of features or observables at a given LoA. (Floridi and Sanders 2004)

Floridi and Sanders further explain that moral agenthood depends on a Level of Abstraction and the guidelines for agenthood are: Interactivity (response to stimulus by change of state, autonomy (ability to change without stimulus) and adaptability (ability to change the transition rules by which state is changed at a given LoA. "An agent is morally good if its actions all respect that threshold; and it is morally evil if some action violates it." (Floridi and Sanders 2004, 378)

It was first Norbert Weiner "The Human Use of Human Beings: Cybernetics and Society", who noticed that the future advance of such communications between man and machines, between machines and man, and between machine and machine, are destined to play an ever increasing part in the future development of messages and communication facilities. (Wiener 1988, 268) Weiner saw that if his vision of cybernetics was realized, there would be great moral concerns raised by such machines which he outlined in this book.

"Cybernetics thinking has influenced contemporary research and development in artificial intelligence and efforts to create autonomous agents (human enhanced, artificial agents) that can self-organize, produce, and reproduce themselves." (Floridi and Sanders 2004)

1.1.4. The concept of cognitive assemblages

N. Katherine Hayle's concept of cognitive assemblages sheds light on the interactions between human and technical systems that enables us to understand more clearly the political, cultural and ethical stakes of living in contemporary developed societies. In her book, Kathrine Hayles, "Unthought; The power of the cognitive nonconscious" looks into and questions perspectives that offer frameworks that are strong enough to "accommodate the exponentially expanding systems of technical cognitions and yet nuanced enough to capture their complex interactions with human cultural and social systems." (Hayles 2017, 23) She also investigates how these patterns present new opportunities and challenges for humanities. Hayles uses the term 'cognitive assemblage' to "describe the complex interactions between humans and non-human cognizers and their abilities to enlist material forces." A cognitive assemblage emphasizes the "flow of information through a system and the choices and decisions that create, modify and interpret the flow." She adds, "While a cognitive assemblage may include material agents and forces, it is the cognizers within the assemblage that enlist these affordances and direct their powers to act in complex situations." (Hayles 2017, 115)

Hayles defines cognition as "a process that interprets information within contexts that connect it with meaning. The meaning of information is given by the processes that interpret it." She explains that "in automated technical systems, non-conscious cognitions are increasingly embedded in complex systems in which low level interpretative processes are connected to a wide variety of sensors." (Hayles 2012, 150) as quoted in (Hayles 2017, 23).

Hayles has developed a framework that challenges the traditional concern in humanities that is concerned with meanings relevant to humans in human dominated contexts. "The framework developed challenges that orientation, insisting cognitive processes happen within a broad spectrum of possibilities that include non-human animals, plants as well as technical systems." (Hayles 2017, 26) The framework sets up the possibility that "cognitive technologies may perform as ethical actors in the assemblages they form with biological life forms, including humans." (Hayles 2017, 27) This signals the possibility of the growing role of Artificial Intelligence in cognitive technologies that maybe designed to be an outright moral defenders and actors.

To highlight that, Hayles gives the example of Peter-Paul Verbeek, who has developed a philosophical basis for thinking about technical systems as moral actors and suggests how to design technologies for moral purposes. “When artifacts embody higher levels of cognition, they can intervene in more significant and visible ways.” (Hayles 2017, 35) Verbeek adds that “moral agency is distributed among humans and non-humans; moral actions and decisions are the products of human-tech associations.” (Verbeek 2011, 53) as quoted in (Hayles 2017, 36). In this sense, the consideration of independent artificial moral agents that can be used as insurgents on social media against violent extremism rhetoric are not that far from reality after all. Hayles argue that we need frameworks that explore the ways in which the terminologies interact with and transform the very terms in which ethical and moral decision are formulated.

To analyze and evaluate the effects of technical cognitive systems, Hayles suggests the need to see the effects from the perspective of what she named “choice II- the consequences of the actions the assemblage performs, instead of remaining within individual-focused frameworks for ethical or moral judgement.” (Hayles 2017, 39) Hayles focuses on non-conscious cognition that she argues “is absolutely essential for higher cognitions, contributing to the very foundations of human cognitive system.” (Hayles 2017, 50). One of the illustrations of her arguments is based on surveying results in cognitive psychology and other fields conducted by (Pawel Lewicki, Thomas Hill, and Maria Gyzewska 1992) regarding the functions and structures of nonconscious cognition in addition to pattern recognition that recall the fact that consciousness is much slower than nonconscious processes. They note “nonconscious information acquisition processes are incomparably faster and structurally more sophisticated.” (Hayles 2017, 51). The power of nonconscious cognition and cognitive assemblages are factors worth noting when considering the possibility of using artificial intelligence for insurgency tasks against radicalization that will lead to violence. A good question perhaps to ask here would be how to integrate agents with non-conscious cognition skills that artificial agents learn to act in a morally desirable way to combat and convince radicalized narratives on social media that can potentially lead to violence?

The idea of cognitive assemblages stresses the importance of considering the role of non-human actors in insurgency efforts against for instance radicalized echo-chambers that promote extreme violence to meet their delusional goals. Hayles gives the example of Pentland’s lab that has developed ‘a computer algorithm that builds on the socio meter’s ability to read the group’s honest signaling. Using this technology they are beginning to build “real-time meeting management tools that help keep groups on track, by providing them with feedback to help

avoid problems like group think and polarization." (Hayles 2017, 125) Researchers in Pentland's lab further note that "important parts of our personal cognitive processes are guided by the network via unconscious and automatic processes such as signaling and imitation. The sociometer developed in Pentland's lab performs in a technical mode operations similar to the human cognitive nonconscious by sensing and processing somatic information to create integrated representations of body states…the human cognitive nonconscious recognizes and interprets behavior, including social signals emanating from others." (Hayles 2017, 126) Such theoretical frameworks are certainly relevant in using artificial moral agents that can process information effectively and be used in the battle against violent extremism on social media.

One of the most applicable recommendations for this research is Hayles proposal of a framework, where she stresses the common assumption of taking 'human cognition' as a whole of cognition or that it is unaffected by the technical cognizers that interpret it. She argues that even though the unique potential of human cognition must be recognized, understanding the situation as a cognitive assemblage brings us close to the reality. "It foregrounds both the interplay between human and technical cognitions and the asymmetric distribution of ethical responsibility in whatever actions are finally taken." (Hayles 2017, 136) "More accurate and encompassing views of how our cognitions enmesh with technical systems and those of other life forms will make better designs, humbler perceptions of human roles in cognitive assemblages, and more life-affirming practices as we move toward a future in which we collectively decide to what extent technical autonomy should and will become intrinsic to human complex systems." (Hayles 2017, 144)

Although many government and non-governmental organizations have similar goals to battle violent extremism that may manifest itself in the form of hate speech or recruiting of youth online by extremist groups, their methods and approaches may vary. Besides, the growing role of non-human actors on social media demands encountering radicalization that has taken into consideration the permeation of human complex systems and cognitive technical systems. Cognitive assemblages as Hayles rightly points out "are inherently political as they are comprised of human-technical interfaces, multiple levels of interpretation with associated choices, and diverse kinds of information flows, they are infused with socio-technological-cultural and economic practices that instantiate and negotiate between different kinds of powers, stakeholders, and modes of cognition." (Hayles 2017, 178) The next section signals issues raised in the case studies of this project.

1.2 Uprising on social media- case of Ethiopia

As a case study, the project investigates the role of social media networks (mainly Facebook) in inciting conflicts and further creating polarization amongst individuals and groups in Ethiopia. Protests and insurgencies prompted highly by the use of Facebook that led to a dynamic change in leadership in Ethiopia; the actors involved and major incidents on social media that had a double-edged effect, in terms of their use as alternative platforms to bring about democratic change in Ethiopia but also the role of Facebook and other social media played in inciting further hatred between diverse groups, particularly divided on ethnic lines, promoted extremist views amongst Ethiopian social media network that activated hatred, violence and radicalization; will be some of the key components of my research.

The project further studies how prevention efforts that make use of several digital tools and technology like the semi-automated identification technology from the Institute for Strategic Dialogue (ISD) can be deployed to resolve and mediate extremist individuals and groups in the Ethiopian social media network. Moreover, it explores the possibility and effectiveness of using artificial moral agents in the defusion of hate speeches, intervene in the recruitment of radical youth by violent groups among Ethiopian social media users, in the aftermath of Ethiopia's recent revolution.

Can non-human actors be effective as identifiable counter insurgency models in the Ethiopian context? In the next chapter, I look closely at the role of Facebook and other networks in leading social movements in Ethiopia and consider how dissident groups used Facebook for civil revolt, while at the same time creating a favorable ground for extremist groups on all corners to operate; often divided on ethnic lines.

In order to analyze the role of social media networks in leading social movements in Ethiopia, it is important to briefly review the historical perspective of the political situation of the country and to understand how social media particularly Facebook became a popular platform for Ethiopian users. What kind of strategies (if any) were used by activists, journalists, prominent politicians etc. to revolt and force the ruling party, the Ethiopian People's Revolutionary Democratic Party (EPRDF) to embark on major changes after almost three decades of repression; for example, successfully campaigning to free political prisoners, journalists,

activists etc.) Who are the popular actors involved in making decisions? Can this be regarded as a collective popular movement or was it orchestrated? In what way does the role played by social media in Ethiopia relate to other social media revolutions like the Arab Spring?

This paper is structured in a way that in Chapter Two, I introduce the case study of this project and give the background basis of conflicts on social media networks in Ethiopia before and after reforms. In addition, I review other notable social media driven uprisings like the Arab Spring and compare it with Ethiopian social media activism and activists. Chapter Three explains the methodological approach of the project. The research design and strategy as well as clarification on how I plan to obtain data from the unstructured interviews is conferred. In addition, the legal aspect of the research and challenges are discussed. Chapter Four presents the results of the study categorically explaining findings related to the Ethiopian reform and extremism. The chapter investigates the role of social media in catalyzing ethnic violence. I also identify some causes and drivers of ethnic extremism, where I attempt to investigate the online and offline link. Further, I discuss the emergence and growth of ethnic extremist ideologies. A separate section in this chapter has also been designated to discuss unemployment, social disparity and propaganda as recipes for violent extremism. Chapter Five brings together the core aspects of the project and discusses the incorporation of Artificial Intelligence (AI) in 'moral actions' on social media. The chapter concludes by chronicling the way forward for counterinsurgency efforts against violent extremism.

Chapter Two

2. Case study- Social media networks in Ethiopia before and after the reforms

2.1 The Arab spring & Ethiopia's activism

Social media networks have been prominent in the last decade for their role in leading social movements. The civil revolt in Moldova in 2009 and the unrest in Iran, popularly known as the Iranian Green Movement, that took place the same year are widely recognized as "the first Facebook revolution" and "the first Twitter revolution" respectively (Dabashi 2013) . Similarly, the spread of the protest would not be swift and practical, if not for efficient social media channels used during the Arab Spring phenomena that began late 2010 in response to oppressive regimes and a low standard of living. The event triggered the solidarity of at least eight countries in the region of the Middle East and North Africa and led to the ousting of regimes in the region including Muammar Gaddafi of Libya, Egypt's Hosni Mubarak and Zine Ben Ali of Tunisia.

The emergence of digital networked communications has certainly eased the process of social movements, in terms of developing collective identities with a common ideology to promote a specific agenda or cause. However, the impact and consequences of social media in fundamentally changing power relations in society remains debatable amongst scholars**.** No extensive research has been carried out either on the specific role of non-human actors such as professional trolls and bots on social media in the Arab Spring and other social media movements like the recent Ethiopian uprising. I will later in the chapter give actual examples of how repressive regimes such as Bahrain and Ethiopia have manipulated social networks using automation to push out vast amounts of political content to gain support by spreading misinformation and junk news.

Bart Cammaerts, in an article from a collection of Westminster papers on the role of social media in the Arab uprisings- past and present entitled 'Social Media and Activism' points out that "even sceptics of the potential of social media in altering power relations in society acknowledge the opportunities for disadvantaged groups to represent themselves, communicate independently and organize transnationally. Social media are playing an increasingly constitutive role in organizing social movements and in mobilizing on a global level**."** (Cammaerts 2015)

Considering the deadly conflicts with extremist groups taking control of towns and areas in Libya after the fall of Gaddafi, Egypt's unstable transition to a harsher military regime, and Yemen and Syria on the verge of state collapse, the success of the social-media-driven Arab Spring in terms of answering important demands of protesters that kindled the movement in the first place remains debatable. Only Tunisia moderately succeeded in achieving a partial democratic transition. There is no doubt however that social media have played central roles in igniting the Arab Spring, although impacts were initially underestimated because of low penetration rates of the Internet in the Arab world.

Lina Ben Mhenni, a leading activist in Tunisia, for instance, was amongst highly effective activists in exposing and communicating government abuses using her Facebook, Twitter and blog; where she took photos of protestors killed by police, uncovered and criticized corrupt officials, called on followers to organize and protest. While her updates continued through the revolution despite several attempts by Ben Ali's regime to silence her. Even if her pro-democracy blog site was blocked by the government, she continued to run it through proxy sites.

During the initial phase of the Tunisian revolution, Mhenni's and several other activists' real time updates of protests helped to spread the revolution. This was particularly apparent following the death of Mohammed Bouazizi, a street vendor in Tunisia, a bread-winner of eight, who set himself on fire in front of the provincial headquarters south of the capital Tunis, where he tried to file a complaint after his vendor cart was confiscated by police for not having the appropriate permits.

Bouazizi's actions and graphic images of him in flames went viral on social media, not only angering Tunisia but other countries in the region and is believed to have started the Arab Spring. There have been several reports of self-immolation following Bouazizi's suit including in Egypt, Algeria and Ethiopia.

Ethiopian activists on social media, for example, were quick to compare Bouazizi's self-immolation to the act of Yenesew Gebere, a 28-year-old school teacher in Waka town in the Oromia region, who set himself on fire in November 2011, in front of the local government building, while protesting the mass arrests of youths from his local area. He died in the hospital three days later from his injuries. Following his death, Ethiopian authorities reportedly switched off the telephone network in the area and restricted residents' movements and the incident was not reported by any media except opposition diaspora websites (difficult to access in Ethiopia

at the time), local human rights organizations could no longer operate, their accounts being frozen, and their field investigators harassed. (Jon Abbink 2017, 67)

Even though Yenesew's act went viral on social media and angered many active Ethiopian internet users both in the diaspora and in Ethiopia, it has not provoked the same kind of popular reaction as Bouazizi's martyrdom did in Tunisia, despite the boiling frustrations of Ethiopians, over autocratic rule, ethnic marginalization, high unemployment and corruption. This is in part mainly because of the government's action of controlling and restricting the social media and the internet.

Although acknowledging the reasons for political developments are complex, Terje Skjerdal, in a chapter entitled "Why the Arab Spring Never Came to Ethiopia*"* argues that formal and informal restrictions in Ethiopian media governance became a major impediment for the public's engagement in a movement for political change. This includes the government shutting down social media networks like Facebook partially and completely, especially when an uprising or protest erupts. (Skjerdal 2016)

On the other hand, some researchers have demonstrated the intricate relationship that exists between offline collective action and social media. For instance, Gerbaudo's article entitled "The 'Kill Switch' as 'Suicide Switch': Mobilizing Side Effects of Mubarak's Communication Blackout" stresses the multifaceted and unsure relations that exists between offline collective action and social media. Based on his empirical research conducted with online Egyptian activists focusing on one critical event during the first days of the 2011 Egyptian revolution, the internet blackout imposed by Mubarak's regime failed to achieve its apparent aim. In fact, the regime seemed to obtain the opposite of what it had hoped for as a huge mass of people took to the streets of Cairo and other cities despite the total internet blackout. Gerbaudo argues that the kill switch turned into a "suicide switch" and ending up giving more energy to protesters highlighting the "complex and ambivalent" relation that exists between offline collective action and social media. (Gerbaudo 2013)

While social media have been effective alternative platforms for the voiceless, enabled organizing and spreading information about government abuses, facilitated collective action against repressive regimes, there were several instances whereby governments have co-opted social media strategies to remain and restore their power. Their tactics to silence dissent ranges

from completely shutting down Internet access by cooperating with state owned telecoms to hiring PR companies, who engage in shaming and trolling oppositions, monitoring and spying on prominent opposition activists, disseminating false news as means of propaganda to serve a regime's interests, engaging in high surveillance of targeted oppositions profile pages and numerous other tactics.

Marc Owen Jones, after conducting a 10-month virtual ethnographic study of Bahrain during the uprising in 2011, has examined how the Bahraini regime utilized social media to subjugate dissent and maintain power. Jones underlies initially the revolution was for pro-democracy, thousands of activists among both Sunni and Shia denominations took to the streets of Bahrain, to demand political and social reform. After the brutal crackdown of the government that resulted in the death of up to 76 people, the Bahrain society was rather polarized along a pro-government versus anti-government divide.

(Jones 2013) observed that the Bahraini regime has implemented various social media strategies to stay in power. For example, security forces or PR companies are believed to be behind thousands of bots on social media who engage in trolling to intimidate oppositions, to bully activists and give the illusion of widespread support from the government. There is evidence for instance that a hoax journalist by the name of Liliane Khalil who used blogs, Twitter and email to build up a convincing fake online persona.

Liliane had claimed to be the US editor of a pro-government blog called the Bahrain Independent, an investigation revealed she was a hoax. After she interviewed several activists for instance, convincing them to hear their side of the story, she passed on their personal information to a pro-regime Twitter user- who then broadcast it on Twitter stating that the interviewees were traitors. Liliane Khalil was deployed by a company named Task Consultancy, employed by the government of Bahrain for propaganda, PR, data mining and intelligence gathering. (Jones 2013, 69)

In a similar move, a report by researchers from Citizen Lab at the University of Toronto's Munk School of Global Affair exposed that the Ethiopian government at various times since 2013 has acquired spyware from three software companies, Hacking Team, an Italian company, the Israel-based spyware manufacturer Cyberpit and FinSpy spyware by Finfisher, a company based in Munich, Germany to covertly monitor online activities of targeted dissidents and journalists critical of the policies of the government.

According to the report, more than 40 devices in 20 countries were infected. The researchers uncovered that the spyware used in the attacks called "PC Surveillance System (PSS) used by Cyberpit has features to covert operation, the ability to bypass encryption and the ability to target devices anywhere in the world and the product is marketed to intelligence organizations and law enforcement agencies." (Solomon 2017)

Researchers with Citizen Lab concluded that, "The fact that PSS wound up in the hands of Ethiopian government agencies, which for many years have demonstrably misused spyware to target civil society, raises urgent questions around Cyberpit's corporate social responsibility and due diligence efforts, and the effectiveness of Israel's export controls in preventing human rights abuses." (Solomon 2017)

In both examples it is evident that hegemonic forces utilize advanced technologies and strategies on social media networks to restore their power. Both suppressive regimes of Bahrain and Ethiopia have utilized spyware and other technologies supplied by private foreign companies to silence dissents. This clearly raises human rights concerns when it is targeted at civilians as was the case in both countries. As governments, companies and organizations including violent extremist groups race to achieve all kinds of political, social and economic ends on social media and the internet beyond borders, the question is becoming not about who controls these platforms but rather over their very nature. Marc Owen Jones writes "battles are currently underway over not only who controls its [the internet's] future, but also over its very nature, which in turn will determine whom it most empowers in the long run – and who will be shut out." (Jones 2013, 27)

In the next sections, I review in brief the political, social and economic background of Ethiopia in order to understand better the possibility of designing social bots that can be used to concoct social good, by for instance mediating amongst polarized groups or detecting false news. How did Ethiopian activists on social media managed to succeed despite recurrent internet black outs and surveillance, deployed by the government eventually failed?

2.2. Ethiopian activists on social media

The attempt to bring the Arab Spring to Ethiopia in 2011 using new media technology along the lines of citizen engagement which had been observed in Egypt and Tunisia at the time turned out to be a failure. Scholars and activists predicted the revolution to haunt Ethiopia, after observing growing support amongst parts of the public ahead of the announced rallies in Addis Ababa in May 2011 protesting against the late Prime Minister of Ethiopia's regime of Meles Zenawi, who was often accused by international human rights groups of stiffing political dissents, detaining and abusing oppositions, journalists, civil rights activists. The government was repeatedly accused of shutting down newspapers, jamming and intervening in satellite channels based abroad and later completely blocking Facebook and other social networks.

Ethiopian population with internet access is estimated at 16.5 million users, which accounts only 15.3% of Ethiopia's 108 million people, according to a Statista estimate in 2016-17. Most Ethiopians use their mobile phones to connect to the internet. In 2017, the number of mobile subscriptions in Ethiopia was at 62.62 million. (Statista Country Report 2019) In cities, internet cafes are everywhere, and laptops are increasingly becoming common. The state monopoly Ethio Telecom remains the only internet service provider (ISP) in the country, unlike most nations; which have multiple ISPs, the Ethiopian government needs to coordinate only with Ethio Telecom to block traffic from certain websites or even shut down access completely.

The Ethiopian government has numerous times cut off the internet completely especially when there are political upheavals, for example, "during antigovernment protests throughout 2017, social media and file-sharing platforms such as Facebook, Twitter, WhatsApp, and Dropbox were repeatedly blocked, including during student protests in December blocks on social media first impacted networks in the Oromia region but later spread to other regions, and eventually manifested in a shutdown of entire internet and mobile networks for days and months at a time." According to Freedom on the Net 2018 report. (Freedom House 2018)

The report particularly underlines the growing role social media and communications platforms began to have in the mobilization of widespread antigovernment protests in the Oromia and Amhara regions since November 2015. Activists have used social media platforms to post information about the demonstrations, and to disseminate news about police brutality as the government cracked down on protesters. They were also able to consistently report the arrests,

trials, and releases of political prisoners and some activists even called for an extensive boycott and protest.

One prominent group 'Zone 9' bloggers, (a name from a state prison in Addis Ababa commonly known as Kaliti maximum security prison, which has eight zones, Zone 9 refers to an outside world they viewed as equally chained by the lack of civil liberties) constituted nine young Ethiopian professionals that launched a blog in 2010 about social, civic issues and later turned to political activism and were critical of the government, urging it to respect the constitution, documenting human rights abuses and violations of law by state actors and reported largely on mistreatment of journalists and citizens as well as visiting political prisoners and published messages from them.

All nine of the social media activist group zone 9 were charged with terrorism and having links with an outlawed US based opposition group Ginbot 7, for allegedly planning attacks, and charged for using basic online encryption tools that journalists use routinely to protect their sources, for receiving digital security training from the Tactical Technology Collective/ Front Line Defenders Security in a Box program and terrorism charges that allege they have created a serious risk to the safety of the public. (BBC News 2014)

All the zone 9 members were detained in April 2014 were charged with the much-criticized Antiterrorism proclamation of 2009 and appeared at a court in Addis Ababa. Advocacy organizations and activists across the world have organized campaigns to call attention to the case of zone 9 bloggers. The Committee to Protect Journalists (CPJ) for instance called for the group's immediate release and stated that they had just been doing their jobs. "Expressing critical views is not a terrorist act. Once again, the Ethiopian government is misusing anti-terrorism legislation to suppress political dissent and intimidate journalists," CPJ said in a statement. (Committee to Protect Journalists (CPJ) 2016)

The group's arrest undeniably made Zone 9 more popular and significant. It has encouraged other digital activists to follow suit and be active on social media. There were growing number of activists on social media platforms that have begun exposing human rights abuses and urging authorities to take action. For example, a Twitter campaign in the summer of 2014 using the hashtag #FreeZone9Bloggers as well as the displayed pictures on Tumblr in support of the jailed bloggers were amongst the very popular campaigns. In July 2015, the charges of three of the

bloggers were dropped and they were released from prison and the rest were released in October 2015 with all charges acquitted. (Al Jazeera 2015)

Another prominent activist who has had a major impact in the Ethiopia social media hemisphere is the controversial figure, Jawar Mohammed, 33, a founder of the Oromo Media Network based in Minnesota, USA. Mohammed stands out as a central player in the Oromo protests. He had catalyzed the protests using his network that broadcasts via satellite and effectively using social media, mainly Facebook and Twitter. He has over 1.7 million followers on Facebook, the highest amongst Ethiopian activists.

Mohammed used these mediums to help in orchestrating demonstrations and broadcast indisputable proof of the government's abuses to millions of followers. The *Qeerros,* (young men mostly bachelors) in thousands follow Mohammed's post attentively. Beginning the fall of 2015, for example, thousands of protesters mostly *Qeerros* organized boycotts and set up roadblocks, paralyzing commerce. Mohammed's Facebook activity was always on the limelight, it even once caught the attention of Facebook CEO, Mark Zuckerberg, as his Valentine's Day photo with his daughter was crammed by comments after supporters of Mohammed were angered following his account was temporarily blocked by Facebook, who later apologized in a statement that spam-detecting systems incorrectly blocked Mohammed's account and removed the block. (Shinal 2018)

Fig. 1. A screen shot of some comments from Ethiopian users under Valentine's Day photo of Zuckerberg and his daughter urging to unblock Jawar Mohammed

Mohammed's block sparked so much anger from his supporters and himself, particularly because the protests were at a decisive stage. In mid-February, his account was blocked two days before Prime Minister Hailemariam Desalegn announced his resignation to calm the turmoil. Earlier that week, he has posted images and videos of politically riotous events in Ethiopia. A day before his account was blocked, he has shared a link to a news report with the

headline, “Ethiopia: Oromia state rocked by protests and killings amid a 3-day market boycott.” And shared a picture of empty streets in the Eastern city of Harar during a boycott and protest.

Mohammed’s incident is a reminder that other corporate actors in this case Facebook play a huge role in conflicts. These corporate actors that can decide as to whether certain accounts should be blocked or stay open have significant impacts on political and protest movements.

Fig 2. Mohammed was appalled by Facebook’s block, he twitted asking FB if they have chosen to side the Ethiopian authoritarian regime

Speaking on Aljazeera TV, Mohammed says people from all over the world snap a picture, record videos and send it to us. “Through Facebook or WhatsApp, we take that, we verify it, we edit it and we air it back to them. You cannot imagine this revolution, this change without social media.” (Al Jazeera 2018).

However, Mohammed’s posts are at times provoking and unverified reports that can potentially escalate ethnic tension. (Example of how Mohammed’s posts affect actions on the ground is further discussed in Chapter Five).

The repressive government of Ethiopia is not the only enemy of Mohammed, who has accused him of terrorism. Critics wondered if the ethnically driven political change that he aspired for has gotten out of control, creating extremists based on ethnicity. There are many who consider Mohammed as a radicalized violent person with the intention to go as far as having a separate Oromo state. This was particularly noted as he returned home in August 2018, thousands of people lined the streets to welcome him, while across the country the *Qeerros*, clashed violently with other ethnic groups and the violence has displaced thousands of people. In Chapter Five, I will consider more into narratives of extremists based on ethnicity amongst Ethiopian social media users to identify challenges facing the society after the reforms, that may perhaps serve

as an ingredient to the possible design of social bots in counter radicalization efforts and to mediate between conflicts.

Activists in the Ethiopian social media spheres are not obviously limited to opposition. There are a good number of hardcore supporters of the oppressive regime of Ethiopia that have been tirelessly battling on social media, deploying various tactics to shame opposition and disseminate government's propaganda so that the regime remains in power. A Facebook leaked document from my informants alleged that some TPLF activists were receiving huge sums of money from the party to co-opt social media strategies during the tense period of the revolution, although it was difficult to get tangible, independent verifications, there were clearly signs of deliberate propaganda campaigns by pro-TPLF campaigns. Such TPLF activists are popularly known amongst Ethiopian social media users as '*Digital Weyane'* to refer to the rebellion that fought against the previous communist regime.

Fig 3. Depicts a meme against Digital Weyane groups by listing and exposing them, a part of naming and shaming strategy on social media.

Such memes used by activists that lists the members of *Digital Weyane* a group allegedly supported financially to work against Abiy Ahmed's administration. The meme shows the picture of Sebehat Nega, one of the founders and ideologue of Tigray People's Liberation Front (TPLF), the group that has seen the domination in central government declined after the arrival

of Abiy from the Oromo Democratic Party (ODP). The meme also lists both in Amharic and English, the pages it considers are extremists, financed by TPLF to create polarization amongst groups that may lead to chaos on social media.

In a surprisingly similar strategy to the Bahrain regime, supporters that have accounts or perhaps bots that are anonymous, most of them have very few followers and just as Bahraini supporters use symbols on their profile picture that show their support like one of the royal family profile pictures or in the case of Ethiopia, pictures of the Tigray regional flag or the president of the regional state. As Jones rightly noted in the case of Bahrain, few people who engage in trolling reveal their identity, often making it very difficult to distinguish them from bots. This was also the case amongst TPLF activists on Facebook.

2.3 Overview of how Ethiopia got here

Located in the horn of Africa, Ethiopia is the second most populated nation in Africa and is home to 108 million (2018) inhabitants amongst which is estimated more than 80 different ethnic groups, speaking more than 80 different languages and 200 dialects throughout the country. The Oromo constitute the largest ethnic group at 34.4% of the country's population followed by the Amhara who account for 27 % of the population and other major ethnic groups include the Somali (6.2%), Tigray (6.1%), Sidama (4%) and Gurage (2.5%). (World Population Review 2019)

The official language of Ethiopia is Amharic although regional states use their own respective languages for office use and in schools. The EPRDF government was formed after the military junta communist regime of Mengistu Hailemariam was ousted in 1991 by the Tigray's People's Liberation Front (TPLF), a rebel group that has fought against the regime for 30 years and was helped by the Eritrean People's Liberation Front (EPLF) to topple Mengistu and preceded with the succession of Eritrea as an independent nation.

After the end of 17 years brutal military junta dictatorship rule of Mengistu Hailemariam's Derge communist regime, many Ethiopians had high hopes for their lives to change, as they aspired for freedom, democracy and economic well-being, as the country remained one of the poorest countries in the world. Following a transitional government in 1995, the Ethiopian

People's Democratic Front (EPRDF) was restructured as a party- constituting itself, the Tigray's People's Liberation Front (TPLF) and other affiliated parties: the Oromo Democratic Party (ODP), Amhara Democratic Party (ADP) and Southern People's Democratic Party (SPDP). A state structure was proposed, and the country was formed as the Federal Democratic Republic of Ethiopia that has a parliamentarian form of government comprising nine regional states and two City Administrations (Addis Ababa and Dire Dawa; the largest and second largest cities by population, are responsible to the Federal Government and are not part of any of the nine regions).

Opposition parties emerged and the private press was relatively open, and critics were relatively tolerated during the early year's rule of Meles Zenawi. He has emerged as chair of the TPLF and Prime Minster of Ethiopia and a strong man figure in the country's politics. Meles consolidated power around his minority ethnic group, the Tigreyans, and concreted political order. Later on, in his years began taking harsh measures against political opponents and the media.

Such harsh measures were especially notable following the contested results of the May 2005 elections, despite international observer groups like the European Commission Observers' Group deeming the election to carry irregularities, marred by the intimidation of opponents and critics. Prime Minister Meles arbitrarily detained several opposition groups and journalists that were accused of violating the constitution, with charges that ranged from treason to overthrowing the government by force. Many international human rights groups and oppositions denounced the arrests, saying they were politically motivated and demanded a reform in the corrupt justice system of the state. During unrests in relation to the May 2005 elections, 196 people lost their lives as protesters were killed by special soldiers assigned by the prime minister. An independent commission established to study the killings later accused the government of Meles Zenawi for using excessive force against demonstrators. (BBC 2006)

J. Abbink in an article entitled "Ethnicity and Conflict Generation in Ethiopia: Some Problems and Prospects of Ethno-Regional Federalism" contends that the post 1991 regime in Ethiopia, "despite its promise and claims to bring solutions, has been less successful than expected in managing ethnic tensions in the country, and has basically only decentralized the problems by defining the sources of conflict to be on the local and not national level." He added, "The federal state is all-powerful, retaining political control and financial-economic resources at the center, but declines responsibility for the emergence, or even production of local conflicts

between ethnic communities on the regional or local levels, of which there has been a dozen since the early 1990's." (J Abbink 2006, 390)

EPRDF's politics that promised the respect of human rights, equality, justice for all ethnicities and a fair sharing of resources for all nations on paper failed to materialize for majority of Ethiopians on the ground. The proposed Federalism structure of government that is based on ethnicity was believed to be an important recipe to maintain stability, equality for all ethnicities and economic growth, after years of implementation has rather given way to establishing political views based solely on ethnic affiliation; as opposed to rational thinking and ideology-based politics, further opening the way for polarization of views, growing hatred amongst different ethnic groups that have otherwise lived peacefully for many years.

Despite some instabilities, growing unpopularity amongst the public and international human rights activists' accusation of the government in terms of its human rights violation, Prime Minister Meles Zenawi continued his reign until his sudden death due to illness in 2012.

The International Monetary Fund (IMF) has commended the double-digit growth Ethiopia has recorded in the past years and has proclaimed the country as one of the fastest growing economies in the world. Oppositions and observers question the economic growth that has barely touched the lives of millions of poor Ethiopians. They also dispute the China-style growth model the EPRDF adopted that claims to have achieved double-digit growth for many years while at the same time growing highly intolerant to dissents and oppressive for majority of Ethiopians, favoring a handful on ethnic lines and affiliation.

Following the death of Meles Zenawi, who ruled the country for over two decades, his then deputy Hailemariam Desalegn took power but rather refrained from making any changes mainly due to allegedly pressured from the TPLF wings, although some contain he lacked a commitment for change and chose not to answer to the demands of the public at large but instead wanted to continue the status quo of his former boss Meles.

Resentment and frustrations continued growing over the maladministration of the ruling EPRDF party. The Irreecha tragedy is amongst notable incidents that haunted the administration and the role Ethiopian social media network played to protest against the act that further escalated to other parts of the Oromia region and later on to Amhara regional state as well as to other parts of the country was highly significant.

A Thanksgiving holiday of the Oromo people that is celebrated every year to mark the end of the winter rainy season of June to September, the Irreecha celebration on October 2, 2016, where an estimated two million people were in attendance turned into a nightmare. According to Human Rights Watch 2017 report, scores of people, possibly hundreds, died at the Irreecha cultural festival, following a stampede triggered by security forces' use of teargas and discharge of firearms in response to an increasingly restive crowd in Bishoftu (Debere Zeiet), Oromia in the outskirts of the capital city, Addis Ababa. The report further stated that some died after falling into a deep open trench, others drowned in the nearby lake while fleeing security forces and quoted witnesses that others were shot by security forces. Moreover, according to the report there is no independent and credible determination of the death toll. Opposition groups estimate nearly 700 died, while the government claimed the deaths of 55 people that fled after police fired tear gas and shots in the air to disperse anti-government protesters. It denied allegations that live fire shots were aimed at protesters and federal security forces were not involved but attempt to contain the protest was carried out by the regional police. (Human Rights Watch (HRW) 2017)

Jawar Mohammed and other social media activists exposed several videos recorded at the scene that reveal numerous gun shots that could clearly be heard as crowds flee. As opposed to government officials claim that security forces were unarmed and there was no live ammunition at Irreecha, activists on Facebook posted photos clearly showing heavily armed security forces at the event and footage of the accounts of several witnesses of gun fire and bullet wounds. Despite the evidence and a United Nations request to send in a team of observers to investigate what happened, the Ethiopian government denied security forces were systematic and that it does not need outside interference to investigate what locally happened.

The Irreecha incident further angered the Oromo people, the single largest ethnic group in Ethiopia that has always protested the systematic marginalization of EPRDF that they have accused of being dominated by the Tigreyans. The protest spread fast to other towns in Oromia regional state mainly by the means of social media and foreign based diaspora radio stations and satellite television stations. Ethiopian social media activists not only continued to heighten their influence, they began to mobilize and organize in groups on social media. Several campaigns like call for civil disobedience, blocking of roads, staying at home calls, coordinating protests etc. were all organized on Facebook at an astonishing pace. After protests started to spread to other regional states like the Amhara, that jointly together with the Oromo ethnic

group constitute more than 50 percent of the population, a nationwide demonstration started to get momentum.

One critical turning point in 2015 was the large-scale protest that erupted against the government's plan to extend the capital city, Addis Ababa's borders into Oromia territory. Numerous people were injured or killed, and the protests gathered a nationwide demand for change. The government responded by its common tactic of restricting internet access when protest breaks out. The EPRDF justifies such actions as a response to misinformation, rumors and unverified reports that flood social media when violence erupts. However, opposition activists argue blocking internet access makes it even more difficult for citizens to assemble peacefully or follow up on what's going on the ground.

Although the actual Arab Spring did not get the expected momentum in Ethiopia seven years ago, a sort of revolution similar to the Arab Spring seem to haunt Ethiopia. Aided highly by social media networks most notably Facebook, the hard-fought political activism on social media paid off, as the country embarked on major changes that would pave the way for democracy like releasing opposition politicians, activists and journalists, unlocking critical websites, inviting outlawed political groups based outside the country, like Ginbot 7 and Oromo Liberation Front (OLF) (Once labeled by the Ethiopian government as terrorist organizations).

On 2 April 2018, Abiy Ahmed becomes prime minister of Ethiopia after the unexpected resignation of Hailemariam Desalegn. The 42 year-old who was first elected to chair EPRDF's constituent party the Oromo Democratic Party (ODP), just a few months after taking office ordered the release of thousands of prisoners, lifted the state of emergency, allowed dissidents to return home and unblocked hundreds of websites and TV channels, as well as ending the state of war with Eritrea of two decades and has normalized relations with the long-time foe. Abiy Ahmed has introduced several reforms that were unthinkable not so long ago.

Despite unprecedented changes however, the social media network is still filled with hate rhetoric on ethnic lines or amongst different extremist political group supporters. On a televised interview with the state television ETV, Prime Minister Abiy condemned the irresponsible use of Facebook by extremist groups that resist the change the country is embarking on.

He said one of the problems in the Ethiopian media atmosphere is not clearly distinguishing activism and journalism. Abiy added there is a need for self-regulation, peer regulation and institutional regulation to have responsible discourses on social media platforms. Speaking to the media in what was considered his first press conference after he took office, the prime

minister has classified the media in the Ethiopian situation which he described falls in four different categories. He named the first 'attack dog' a sort of social media attack that tries to find the faults of governments by any means possible even fabricating false news to discredit the legitimacy of the government and attacking it no matter what it does. The second one he referred to as a 'lap dog' a media that collaborates with the government to hide the truth about wrong deeds that may be committed and is loyal to government propaganda. The third one, 'watch dog' which is his ideal that questions and reports by exposing on wrong doings and injustices, while the fourth one is the 'guide dog' that shows directions and forwards solutions to society's problems.

According to Abiy, "even though, false news, hate speech, fake accounts disseminating hate are worldwide problems: we have documented that there are organized groups on social media platforms that run more than ten accounts on social media to disseminate hate and violence that are deliberately working to create chaos in the country. There is a need to draft new hate speech laws to combat this," Abiy said. (Ethiopia News Agency (ENA) 2018)

The new initiative to draft new laws was announced in November 2018 and the draft bill is yet to be voted on. The draft new law imposes fines or jail sentences for activists and press organizations deemed to be inciting violence or spreading hatred online. *Addis Standard* reported that "Ethiopia is struggling from the surge of hate speech and fake news in its limited cyberspace. Therefore, the office of the Attorney general is preparing a draft bill aiming to curb hate speech and bring accountability towards public speeches and every other discourse, which is deemed to ignite hate and ethnic tensions in the country." (Tsegaye 2018)

This signals the government's position of heavily blaming the rise of social media platforms such as Twitter and Facebook for the recent rises in ethnic tensions over the course of the last decade.

However, in response to the new draft law, groups such as Human Rights Watch (HRW) caution that "any law that limits freedom of expression by punishing hate speech must be narrowly drawn and enforced with restraint, so that it only targets speech that is likely to incite imminent violence or discrimination that cannot be prevented through other means, " stated HRW under an article: "Tackling Hate Speech in Ethiopia :Criminalizing Speech Won't Solve Problem", who further warned that "many governments have tried and failed to strike the right balance, and Ethiopia's own track record offers reason for alarm. In the past, the Ethiopian government

has used vague legal definitions including in its anti-terrorism law to crack down on peaceful expressions of dissent." (Horne 2018)

The next section looks into the present situation of Ethiopia's social media after significant progress has been made on media freedom in the country. The country has in one year gone from being one of the leading jailors of journalists in Africa to having no journalists in jail for the first time since 2004. Despite the progress, "there is still hate speech on social media especially Facebook, is a serious and growing problem although the government's proposed hate speech law raises concerns it may be used to stifle legitimate expressions of dissent." (Committee to Protect Journalists (CPJ) 2019)

2.4. Ethiopia's social media atmosphere after the reforms

After many positive human rights reforms and a renewed sense of optimism following several years of protests, the transformation as Florian Bieber, a professor at Austria's University of Graz has rightly predicted, has seen ethnic conflicts increase in intensity and number, "both as a result of a backlash by conservative forces rejecting the rapid reforms or due to the sudden liberalization of the public space." (Goshu and Bieber 2019)

As a preliminary research for this project I have browsed Ethiopian social media networks as a user, focusing on groups I have categorized are radicalized with a special focus on activists from the three main ethnicities: the Oromos, Amharas and the Tigreans (as the elites from these groups dominate Ethiopian social media discourses.) In addition, I have also followed activists from Sidamo and Walayeta ethnic groups from South Ethiopia as there were clashes between the two groups. I was able to notice a clear change in patterns of their behavior, in terms of for instance those that I have considered modest, rational and democratic turned completely ethnocentric, manifesting and exchanging respect less and dehumanizing insults in Facebook random exchanges: many activists that I have considered radicalized have slightly become moderate in their discourses after the reform.

A local newspaper, *Addis Fortune* in its editorial entitled "Protect Free Speech, Tolerate hate Speech," argues in contemporary Ethiopia, "Just because the political and media landscape has been liberalized, citizens and the political elite do not suddenly adopt rational discourse and a sober exchange of ideas." Adding, "that is why it should not come as a surprise to anyone that much of the bias, anger and absolutist views nurtured in the underground are now emerging and

spilling into the mainstream conversation." (Addis Fortune Editorial 2019) The observation indicates the impacts of the unfortunate outcome of years of limitations imposed on free speech by the state.

Even though many decisive positive political changes that have taken place in the country in 2018, a record high figure worldwide of about 2.9 million new displacements associated with conflict were recorded in Ethiopia. The Internal Displacement Monitoring Center (IDMC) in a report states, "Despite many important and positive political changes that took place in the country in 2018, old conflicts became more entrenched and new conflicts escalated along various state borders." The report further stated, "After two decades of relative calm, the most significant displacement was triggered by inter-communal violence between the Guji and Gedeo ethnic groups that erupted in April and again in June 2018 in the West Guji zone of Oromia and the Gedeo Zone of the Southern state." (IDMC report 2019)

On a positive note, some Ethiopian activists based in the diaspora have used social media and the GoFundMe website to raise funds for the victims displaced communities. For instance, a well-known artist and activist, Tamagne Beyene, who has managed to raise over 1 million USD for the victims and has donated through the organization he leads Global Alliance for Ethiopia, a non-profit organization based in the US.

In general, however, Ethiopian social media platforms have continued to be and perhaps have increasingly become dominated by false rumors, sound bites of audios, videos edited to meet political goals and used out of context and photo shops of several deceiving pictures circulate around the social media. At times the impacts of such activities on social media have directly affected people's lives on the ground.

One such case was an incident related to false rumors that went viral on Facebook that have actually led to two medical researchers in Ethiopia being killed by an angry mob in October 2018 and one person survived but severely beaten with a life changing injury. The mob was convinced by rumors on Facebook that the researchers were there to poison the children. The three research scientists from Addis Ababa travelled to Gonji in the Amhara region to investigate intestinal worms and the eye disease trachoma at a local school, reported *BBC News* in a program entitled, "The rumor that led to medical researchers in Ethiopia being killed by a mob." (Irungu and Berhanu 2019) This exemplifies the impacts that false news circulated on social media can have on the ground. False news on social media has the potential to lead radicalized groups cater into violence, causing loss of lives and property. I will give more

related examples in Chapter Five, when I discuss the use of AI in detecting false news on social media.

As I was interviewing one of my informants for this project, an interesting but shocking news story we both could not avoid broke out. On a Saturday evening in June 2019, in a reportedly failed coup attempt, the Amhara regional state president, Ambachew Mekonnen and his two colleagues were shot and killed in the city of Bahir Dar while they were in a meeting. Prime Minister Abiy Ahmed appeared on national television on the same evening on a military uniform and announced that the chief of security of the region, Asamnew Tsigue, was behind the killings and accused him of using mercenaries to orchestrate the attack. Two days later he was shot and killed by police forces as he attempted to flee. Within a few hours after the assassination in Bahir Dar, more shocking news emerged from Addis Ababa that Chief of Staff General Seare Mekonnen, along with his former colleague, was shot and killed by his bodyguard at his home. There was no enough evidence to confirm that the two incidences were connected. The government in a statement said the two incidents were related but refrained from giving further details. (John and Dean 2019)

As expected, the social media went into speculations and conspiracy theories surfaced about what has actually happened. Commenters ranged from accusing the Prime Minister for the attack to favoring and admiring chief of security Asamnew Tsigue for fighting for the rights of the Amhara people. He was especially favored by youth amongst emerging Amhara nationalists, as historically the majority of Amharic speakers affiliate themselves with Ethiopianism. New organizations to promote Amhara nationalism like the National Movement of Amhara (NaMA) have emerged and is recently becoming influential in the country's politics. (NaMA's radical tendencies are investigated further in Chapter four.)

In the wake of the attack, the government shut down the internet. Ethiopia has experienced internet shutdowns since 2015 but it came as a shock to many Ethiopians to wake up to an online blackout since many had high hopes after the reforms that the practice of internet cuts will end. Even though the failed coup is seen as the biggest challenge since Abiy started political and economic reforms in April 2018, critics were against the shutdowns. People had to shift to local media outlets to follow up on events. The inconsistency of interviews by the police and the Prime Minister's office public relations on national media created growing suspicions amongst the public toward the government that began speculating further about the assassinations. There have been claims and counter claims by authorities after the killings in their meeting with the media.

Different radicalized actors including Ethno-nationalists, opposition parties on Facebook have used this blunder to their advantages by providing narratives that support conspiracy theories that were viral on social media. They mainly accused Prime Minister Abiy for orchestrating the attacks to weaken the Amhara elites and called on supporters through social media to protest and denounce the action. Some Amhara extremist activists for instance have used a simple FaceApp to manipulate a photo that made the Prime Minister appear smiling at the funeral service of the assassinated officials. Reading the comments under the post, many naive followers that are not aware of such photo altering techniques have accepted the picture as authentic. This has caused furor amongst youth in the Amhara region. Such posts however are rarely verified by other actors and is one of the methods used by some radicalized Ethiopian activists to wield power through social media.

Amongst the critics of the shutdown of the internet in Ethiopia after the reform is Alp Toker, executive director of NetBlocks, a nonprofit organization that monitors internet censorship. Toker condemned the decision to shut down the internet. “At a time when the nation should be reflecting on the weekend’s events and coming to terms with the loss of life, they are instead denied information and a voice. The loss of dignity and symbolism couldn’t be more striking,” he said, further arguing, “switching off access will only delay and radicalize critical voices as the government is likely to realize when the shutdown ends and Ethiopia’s internet users start coming back online.” (Mbah 2019)

Many communication scholars argue that a more pragmatic and perhaps long-term solution to combating hate speech is not to shut of the internet but to address hate speech as a more of a symptom rather than a problem. For instance, an editorial of *Addis Fortune* argues that in Ethiopia “political discourses in the past were moderated by a generally conservative and inward-looking public sentiment and the intolerance of successive regimes to any radical views. While the state’s repressive measures were counterproductive in addressing the sociopolitical challenges it faced, it managed to create the illusion of stability.” (Addis Fortune Editorial 2019). This highlights repressive measures such as the actions of the Ethiopian authorities to shut off the internet may only help temporarily to calm conflicts in the short term. However, repression is by no means an acceptable method in the long term to combat hate speech or radicalized opinions that may lead to violence.

Abiy Ahmed’s government has begun to be criticized by groups that say he is lenient on groups and individuals whom they consider radicalized and willing to use any means including violence to achieve their goals. For example, the following meme that was circulated by

activists who have become growingly concerned that activists like Jawar Mohammed are highly radicalized and have lead and potentially can lead to conflicts, even though they believe the administration of Abiy has been too tolerant and not taking action.

Figure 4: depicts a meme that shows activist Jawar Mohammed on the back of the Prime Minister to criticize his tolerance for extremism. The comment underneath was ironically made by Journalist Elias Gebere, editor of the Enqu magazine and a member of 'Baladra' a group that claims it has been given the authority to defend the rights of residents of Addis Ababa was arrested on July 4, 2019. This raised questions amongst international human groups like Amnesty International as to whether such arrests are a hugely regressive move that risks rolling back the progress witnessed in 2018 in Ethiopia.

How can we manage to devise a strategy that constitutes the use of artificial moral agents that can systematically learn and adopt the contemporary Ethiopian social media network environment and identify radicalized views, pictures or videos that may promote violence and work as moral agents to defuse hate speeches, expose false news, and mediate between polarized ethnic groups? In order to investigate this, I have interviewed real time users of Ethiopian social media platforms. By identifying central problems associated with radicalization that may lead to violent extremism, the project looks for trends, links, associations of users experience to determine best counter-insurgency plans that work well

against violent extremism on social media. In the next chapter, I will explain the methodology and research processes of this project.

Chapter Three

3. Methodological approach- Processes of the project, methods of data collection etc.

3.1. Research design & strategy

The research strategy was developed with the view of developing a better understanding the uses of social media that can incite hatred and violence amongst Ethiopian social media users as well as to further investigate the conditions that have shaped the relationship between power, politics and social media in Ethiopia. The increasing concerns about uses of social media that can buttress radicalization and incite violent acts has led to increasing demands for research that can expose and monitor these types of online behaviors. The research strategy was designed specifically for the case study of the project, Ethiopian social media users, but such a polarized environment on social media, as I will consider in Chapter Five is not exclusively an Ethiopian affair, as there are many similarities in how social media users tend to become radicalized worldwide.

The sampling of selected five Ethiopian users that represent different groups would perhaps enable us to have a nuanced perspective of political, ethnic and religious divides that has rocked the Ethiopian social networks in recent years, especially in the aftermath of the reforms led by prime minister Abiy Ahmed. Furthermore, the research design based on the informants' interviews explores the possibility and effectiveness of using artificial moral agents to detect and take measures against hate speeches, intervene in the recruitment of radical youth by violent groups among Ethiopian social media users.

3.2. Unstructured interview

The methodical approach selected for this project was to conduct unstructured interviews with five active Ethiopian social media users. This approach has taken into consideration how subpopulations within social media platforms behave differently based on for example, ethnicity, race, sex, age, country, user group members etc. Profound interviews with users and activists allow us to better examine facets of Ethiopian social media users' perceptions, attitudes, behaviors, uses and experiences with social media. It will also serve in identifying trends and actual problems in the Ethiopian social media environment; which may need to be considered in the incorporating of artificial agents like social bots that can be built as models in

counterinsurgency strategies against activities that are destructive to human actors such as dissemination of false news responsible for mob killings or extremist hate speech directed at attacking a specific ethnic or religion that encourages violence and destruction.

If autonomous agents are to engage in social communities such as Ethiopian social media networks, these agents will be expected to follow the community's social and moral norms relevant to the tasks they would be able to perform; which amongst others can possibly be integrated to function as a mediator between polarized groups or individuals on social media. Thus unstructured interviews with informants are integral to identify the norms of the Ethiopian community on social networks, as "different types of technical embodiments will demand different sets of norms". (Created by committees of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2018)

Further, unstructured interviews serve to extract the experiences from actual users and ideal as a qualitative research method to collect data. This is mainly because they allow to create a context in which the researcher's control over the conversation is intended to be minimal and the participants are free to share experiences; providing one with exhaustive information that can later be sorted accordingly. Instead of conducting a one-time face-to-face interview, during this research, informants were contacted several times through Messenger, Viber, Telegram, and WhatsApp. They were encouraged to share interesting materials they came across relevant to the research. My focus was on violent extremist contents, provocative materials, pictures and videos that they thought may incite violence. Particularly, I encouraged participants to share items that are viral on social networks breaking through standards set by social media companies such as Facebook's Community Standards. Sort of moral obligations that users must oblige to that are designed with "a set of rules and orders to combat violence and criminal behavior that apply to all types of content with the view of promoting safety, voice and equity."(Facebook n.d.)

However, the unstructured interviews were not random and non-directive. As much as possible, informants were encouraged to relate to experiences and perspectives that had relevance to the research question. Broad questions that will later allow for a descriptive analysis of users; such as motivation of using social media platforms, identity presentation, the role of social media in interactions, privacy and disclosure of information etc.

Yan Zhang and Barbara M. Wildemuth in a chapter on "Unstructured interviews" point out that unstructured interviews can be useful in studying people's information seeking and use

especially in studies for attempting to "find patterns, generate models, and inform information system design and implementation." (Zhang and M. Wildemuth 2006, 2) Subsequently the project assesses how non-human actors can be effective as identifiable counter insurgency models on social networks, unstructured interview as a methodological approach is effective to examine user's information activities on social media platforms better than other more structured methods due to the informal, 'conversational and non-intrusive' characteristics of the unstructured approach. (Zhang and M. Wildemuth 2006, 3)

3.3. Choice of informants & role of the researcher

Interviewees were selected based on their active engagement in the Ethiopian social networks and their habitual use of social media for at least the last five years. A preliminary study has been carried out as social media activities like posts and comments have been assessed to determine informants represented a diverse group of Ethiopian communities; for instance, in terms of their ethnical and religious background as well as political views. The interviews were conducted in Amharic language, (the official working language in Ethiopia) apart from one of my informants, whose mother-tongue language was Oromo and chose to alternatively speak in English and Amharic. (Zhang and M. Wildemuth 2006) stress the importance of understanding the language and culture of the interviewees; as the primary focus of an unstructured interview is to comprehend the meaning of human experiences from subjects' perspectives. Hence, "unstructured interviews are governed by the cultural conventions of the research setting. This requires that the researcher can understand the interviewees' language and, further, its meanings in the specific cultural context of the research setting." (Zhang and M. Wildemuth 2006, 4)

Follow-up questions and explanations with informants were important because events in Ethiopia have been unfolding at an astonishing pace for the last several months and there is great dynamism in activities amongst Ethiopian social media users. Moreover, my role as a researcher involved following through the subjects' narration and improvising questions based on their reflections. A great deal of effort has been made to strike a balance between consistency and flexibility. Nevertheless, I obviously had some specific goals coming into the interview process, even if it was unstructured. (Corbin and Morse 2003) describe the unstructured interview as "a shared experience in which researchers and interviewees come together to create a context of conversational intimacy in which participants feel comfortable telling their story." (Corbin and Morse 2003, 341)

My role as a researcher is perhaps important to mention, as my choice of Ethiopia as a case study is due to my personal involvement in the background of the story. My background as a reporter for three different local newspapers (Capital, The Sub-Saharan Informer, Addis Fortune) weekly newspapers in Ethiopia as well as working as a freelance journalist for the Associated Press and Bloomberg news agencies based in Ethiopia has driven my interest to research the Ethiopian situation further. However due to my personal involvement of the story, I tried to be aware of and to strive to mitigate my own subjectivity in conducting the interviews, as some of the issues raised were sensitive as I was personally subject to harassment and imprisonment by Ethiopian officials for my work as a journalist on multiple occasions.

3.4. The legality of data use

Norwegian center for Data Research (*Norsk Senter for forskningsdata*) (NSD) has assessed that the processing of personal data in this project will comply with data protection legislation. The planned processing of personal data will be in accordance with the principles under the General Data Protection regarding:

-lawfulness, fairness and transparency (art.5.1 a), in that data subjects will receive sufficient information about the processing and will give their consent

-purpose limitation (art.5.1 b) in that personal data will be collected for specified, explicit and legitimate purposes, and will not be processed for new, incompatible purposes

-data minimization (art 5.1 c) in that only personal data which are adequate, relevant and necessary for the purpose of the project will be processed

-storage limitation (art 5.1 d) in that personal data will not be stored longer than necessary to fulfil the project's purpose.

Because the five informants have considerable number of followers on social media and represent perhaps different extreme wings, it is important that their views and experiences are candidly revealed so that data can be used for counter insurgency efforts like mediation between polarized groups by low enforcement agents among Ethiopian social media users.

All participants were provided with information regarding the purpose of the project, why they were being asked to participate, who is responsible for the research project, what participation means and requires, an explanation of how data will be stored and processed as well as what

will happen to the data at the end of the research project along with their specified rights were handed to them prior to the interviews.

3.5. General interview format and main questions raised

In order to categorize issues raised by informants in themes, the leading questions raised in the interviews were given to the informants prior to the interview. However, the interviews were certainly not limited to the written questions, but much more improvised flexible conversations have been carried out. The main questions of the interviews covered five major aspects:

The first one asks interviewees what changes in behavior they have observed (if they have any) among Ethiopian social media users after the current reforms led by Prime Minister Abiy Ahmed is beginning to take effect, as the right to criticize the government on traditional and social media alike will not subject individuals to serving a jail time or abuse by authorities as it previously did. Do you think this has changed the way Ethiopians on social media behave in response to your posts or personal messages you receive from them?

The second set of questions inquires if the interviewee has come across any physical attacks because of his/her posts or interactions on social media? Explain how you deal with trolling and name shaming you face on social media? In your opinion, to what extent do you think your followers and other Ethiopian social media users are likely to consume extremist content? How are people presenting themselves on Facebook? Do you consider some of your followers as being radicalized and can potentially turn violent? If so, explain why and perhaps give an example.

The third set of questions raises the extent to which social media is affecting relationships among individuals and groups in Ethiopia? What do you think is the motivation of using Facebook and other social media for your followers? Social media networks have increasingly become filled with misleading information, false news and hate speech in Ethiopia. Can you elaborate how you verify your posts? How careful are you about sharing material, posting, commenting and replying to posts about ethnic, religious sensitive topics for instance?

The fourth set of question asked if informants can identify any Facebook or other social media groups that intentionally maneuver to spread hate and violence without being blocked. How do you think such actions that lead to violence and destruction should be handled? Have you

experienced or detected bots engaged in trolling or otherwise? How do you deal with these social bots and pseudo accounts?

The last set of questions tries to establish what informants think about counter insurgency strategies against violent extremism on social media specifically designed for Ethiopian users should be adopted? To what extent do you and your followers are aware of the new draft law, which states for instance, whoever intentionally intimidates or threaten another person or his families with serious danger or injury by disseminating any writing, video, audio or any other image is punishable up to 3 years of imprisonment. Do you think such laws are effective in practice?

3.6 Methodological challenges

One of the main challenges was the time-consuming process of arranging the interviews. Process of conducting interviews were at times difficult due to bad internet connection in Ethiopia and subjects were not always available and it was at times problematic to get in touch with them. Each interview was individualized, and the length of each unstructured interview sessions varied making it challenging to determine how much time I needed.

Another challenge of the unstructured interview was to apply the right amount and type of control over the direction and pace of conversation; it was difficult at times to detect whether to let informants follow through so that they would not lose continuity or to stop them when speaking off topic conversation that seemed irrelevant to the research. Moreover, analyzing the collection of data from unstructured interviews are challenging to structure due to the difference in contexts across the multiple interviews. This is mainly because the questions are not standardized but were rather based on the responses of the interviewees, generating different responses. An attempt has been made to systematize the data so that patterns and relations can be observed from the interviews.

In the next chapter, the results of this research project are presented. For the sake of clarity and order, I have categorized topics based on relevance to the research question and my area of research interests, issues informants raised that are central in understanding polarized ethnic and political conflicts in Ethiopian social media platforms. Where it is relevant, I have emphasized topics that need further discussion based on the informants' reports, that are additionally analyzed in chapter five, with an emphasis on how the data obtained from the

interviews relates to and aids in the investigation of how non-human actors such as artificial moral agents might be effective as identifiable counter insurgency models in battling social media radicalization that may lead to violence.

Chapter Four

4. Results, presentation of data, outcome, findings, interpretations

4.1. Reform and extremism

The much-anticipated reform, after about a year and some months, saw along with it a rise in troubling deadly ethnic tensions in different parts of Ethiopia. Observers including my informants are of the opinion that voices and complaints that have been stifled by decades of repression are now surfacing out on social media platforms and local traditional media, especially FM radio stations as well as media houses owned by regional governments. Human rights advocates estimate that some 1,200 people have been killed and some 1.2 million displaced due to ethnicity related conflicts since Abiy Ahmed took power more than a year ago. (Cara, Meseret, and Moulson 2019) Critics including two of my informants are concerned the unrest will grow violent ahead of the 2020 elections.

Ethiopia's awaited reform that was acknowledged worldwide for its positive role for the future of the country and the Horn region in terms of bringing peace and preparing the path for building a democratic nation faces enormous challenges amongst which is growing ethnocentric disputes that are often orchestrated by ethnonational extremists that use different tactics on social media and promote their anger by calling for violence against certain ethnic groups.

One of my informants, **Andualem Sysay**, a freelance journalist, who is also active on social media platforms describes the current escalating contentious ethnic violence in Ethiopia:

> "I can compare the current reform that is led by Prime Minister Abiy Ahmed to the opening of a house door that has been closed for many years. When the door is open, the air containing heavy odor comes out of the room and the bad smell would remain for some time. Long held grievances that were previously very frightening to say in public have begun to emerge on social media platforms. It would have been great if these were grievances that are addressed in a manner that we can learn from the mistakes of our forefathers. But instead, we are witnessing different groups having a rather polarized understanding of history and want to solve their misunderstanding by violence. It is sickening. Activists purposely manipulate history to fit their narratives and create divisions amongst groups. Ethnic fanatics are very quick to take advantage from a divided community that is vulnerable to unsubstantiated propaganda. The

> majority of young Ethiopian social media users are gullible and naïve, lacking a common understanding of history. We are witnessing the increased use of terms like *Gala*, *Neftegna*, *Ye Ken Jib* that characterizes the major ethnic groups in the country namely the latter *Gala* to refer to undermine Oromo ethnic group, *Neftegna* to refer to the Amhara and *Ye Ken Jib* mainly to refer to Tigreans on social media. These demining names have irritated and escalated tensions between different extremist ethnic groups. These racist discourses are not new in Ethiopia but social media especially Facebook has enabled to disseminate such destructive discourses in a scale that is astonishing."

Such views reflect how slurs and labels on ethnic lines have fueled conflicts as well as the impact of historical grievances amongst different groups resulting in the proliferation of violent extremism. Social media has been exploited by extremists to disseminate false propaganda to instigate feud between different ethnic groups by amplifying historical discrepancies that are often distorted.

Some observers compare the current rising Ethiopian ethnic tension and the role of social media to the impact of radio programs such as *Radio Television Libre des Mille Callines* that spread the toxic hatred that fueled the Rwandan genocide. They are of the opinion that social media in the Ethiopian context appears to be just as effective in spreading untruths and spikes of ethnic violence in the country. They call for urgent action from the government, elders, civil society and other concerned groups to step up efforts and work in collaboration to avoid Yugoslavia like disintegration Ethiopia faces at the moment.

The rising autonomy of regional states, the war of words between regional leadership even among constituent parties within the ruling party EPRDF coalition, has further exacerbated feuds between extremist supporters of the Amhara and Tigrean ethnic groups. A case in point, following the failed coup attempt that has left the Amhara regional president and his colleagues assassinated, the Tigrean TPLF party has issued a provoking statement. The TPLF accused the Amhara ADP party, which it blamed for the assassination of chief of staff of the Ethiopian Defense force, General Seare Mekonnen and his longtime friend General Geazi Abera. Both were reportedly shot dead by Mekonnen's body guard at his residence. This happened hours after the regional president and his associates' were killed in Bahir Dar. The TPLF further asked the ADP to apologize for the Ethiopian people and take responsibility for the murders and characterized Amhara people as chauvinists. It also gave an ultimatum for ADP that TPLF is increasingly finding it difficult to work with its sister party ADP. This has obviously angered the ADP leadership, who were still mourning the death of their colleagues and made a statement

in response to TPLF. ADP retorted that the TPLF leadership are exploiting the current situation in the country to cover up the crimes the party members committed for many years. ADP further called on Nations and Nationalities to continue struggling for the reform measures and accused TPLF of plotting against the reform.

Such polarized and extremism narratives, especially when they come from government officials and party leaders, attract social media ethnic hardliner activists. After the reports from the two parties, social media platforms were filled with comments that supported the ADP's rhetoric. The ADP had never before had the tradition of standing up against its sister party the TPLF. Only few commented against the two parties indulging in publicly insulting each other, opening the way for many radicalized followers with the potential to cater to violence.

My second informant, **Endale Assefa**, is an activist who advocates for a reconciled Ethiopia and a nationalist who works for the government. He actively writes articles and comments on his Facebook page. He says the leaders of both the TPLF and ADP parties should have been more responsible and exemplary to the youth.

> "It is really sad and at the same time raises a huge concern about the future of the country. We are witnessing a new trend in the ruling party's coalition member parties as they are publicly insulting and accusing each other often using terms that are dividing for the nation. I have noticed that not only dissatisfied, oppressed individuals radicalize but also officials in government that are highly becoming extremists on ethnic lines. The official war of words between TPLF and ADP is of huge concern in influencing youth on social media to radicalize. I fear such exchange of dialogues are potentially dangerous as they signal the possibility of a civil war. Instead of being visionary leaders by guiding the youth to disengage from violent extremism, they are doing the direct opposite by aspiring vulnerable youth to be extremists. It is utterly irresponsible for a leadership that is looking forward to a new Ethiopia, one that is reconciled and conflict free."

Assefa's observation relates to Schmid's radicalisation theory, where he notes that in some political circumstances not only non-state actors but state actors like governments can radicalize too. In the case of the growing feud between the TPLF and ADP restoring to radicalized ethnic slurs reminds us of the considerable impact such actors play to influence their followers and other social media users to be radicalized and eventually engage in violent extremism.

Scholars and three of my informants believe holding elections in 2020, before the grievances and reasons for the rise of ethnic politics that are for instance the need for constitutional reform and the strengthening of independent institutions is achieved, they predict the elections will have a disastrous outcome. As one informant said holding elections will only be a choice between an ethnic identity and multiculturalism.

Kebour Ghenna, a public scholar, who actively participates in public debates and regularly writes on his Facebook page warned that (Ghenna 2019b) "this will be an election where our chronic lack of shared vision as a country will clearly be manifested. Without a shared vision, the tendency is for citizens to devalue one another as they glorify themselves and their ethnic identity believing that they are the best thing to have happened to their country. Such selfish greed hardly allows for time to think of Ethiopia in real terms as a collective treasure to be valued and protected by its inhabitant." He added that "this is an election devoid of ideas in dealing with the rise of ethnic hatreds. PM Abiy knows it, his government has yet to define Ethiopia's national order. Reflect on whether centralism or decentralism will be the governing political strand of the country, develop a sound architecture of political institutions and rules, consensus with each regional state having a veto power on critical decisions." (Ghenna 2019b)

The political standoffs between constituent parties of the ruling coalition, opening up of the political space after decades of repression, manipulative activists that take advantage of distorted historic narratives are central points that we need to take special attention to in future counter-insurgency strategies against violent extremism on social media in the Ethiopian context.

As signaled in Chapter Two, understanding the role of social media and how it relates to the actions of violent extremism on the ground requires extensive research. However, my subjects gave me some insights to how their experiences on social media and actual events on the ground interrelate. In the next section I investigate how radicalized individuals and groups use social media to catalyze ethnic violence.

4.2 Role of social media in catalyzing ethnic violence

There is no doubt that social media, especially Facebook, has been a popular platform for Ethiopian radicalized groups and individuals. As signaled in the examples in Chapter Two, social media is strongly tied to ethnic violence on the ground. That is why even after the reform

took place, the government shut down the internet completely in the aftermath of mass arrests following the recent failed coup attempt. The government of Ethiopia has publicly accused social media of being used to incite and exacerbate ethnic violence in the country. The Ethiopian diaspora community believed to exceed two million in number were divided in opinion, as there were some who supported the action of shutting down the internet to prevent the escalation of violence. Others argue that shutting down the internet only prevents the public from following up on events. They see it as a major setback for the government of EPRDF with the leadership of Abiy Ahmed that had promised amongst others not to shut off the public from getting any information they want from the internet and any private media sources. They are saddened that the only solution the EPRDF knows when conflict arises is to shut off the internet instead of addressing the root causes of the conflicts.

My third informant **Tamiru Addis**, who is a civil engineer by profession, supports National Movement of Amhara (NAMA), (a new party that other informants described as an 'extremist Ethno-Amhara group') for instance believes the reform is hijacked by hard core Oromo extremists, who are doing the same thing the TPLF did as they feel it is their ethnic group's turn to lead the country and consider it as their chance to abuse and torture as a sort of payback. He says, it is not surprising that the government still shuts down the internet when conflict arises, people have forgotten the EPRDF is still the same coalition and there are still some old faces in government, who have not accounted for their crimes in the past decades. Addis says that:

> "The majority of ODP party cadres want to promote the interest of extremist Oromo individuals and groups, as the likes of Jawar Mohammed appear more powerful and influence the officials in making important decisions. Jawar at one point has publicly dared to say there are two governments in Ethiopia, the Qeerro government and Abiy's. Such statements clearly indicate how Abiy's party ODP and the federal government is under immense influence of extremists. I have lost all the trust that I had for this government. I have noticed that when Jawar posts something on his Facebook page, government officials even participate and like his posts. For example, ODP party's office chief, Addisu Arega is an adamant supporter of Jawar. I was first very delighted and full of hope like the majority of Ethiopians, when Abiy Ahmed came to power about a year ago. His ideal of love, unity and forgiveness, as well as his Ethiopian nationalism rhetoric was very appealing. Plus, Abiy is a gifted orator. But my trust and hope for this government is diminishing by the day. It appears we are at the risk of losing our country

(Ethiopia) because he is too lenient on the extremists both within and outside his party. Abiy's catch word '*Medemer*' that is supposed to be uniting the country doesn't seem to be working."

Medemer is an Amharic term Abiy Ahmed has coined to describe his philosophy and recently published a book entitled *Medemer*) in three languages namely Amharic, English and Oromiffa. He refers to his ideology as 'synergy' in English where he describes a reform strategy that are centered on unity and three independent pillars, namely, building a vibrant democracy, economic vitality and regional integration and openness to the world.

Ethiopians who share Addis's opinion are often reflected on social media and those who are losing trust in the incumbent government are not few in number. They believe the government of Abiy Ahmed favors the Oromo ethnic group, whom they accuse of playing double-standards; that's harsh on Ethno-Amhara extremists, while being lenient to take actions on Ethno-Oromo extremists. They often refer to Oromo activist Jawar Mohammed whom they accuse of orchestrating attacks through his Facebook page against other ethnic groups. They are of the opinion that the TPLF led EPRDF that only worked to appease for Tigrean ethnic group has only been substituted by the Oromos. However, such views are far from being universal as there are others who believe the reform has its flows but they trust in the positive intention and vision of Prime Minister Abiy Ahmed.

My fourth informant, **Getachew Nigatu**, an active Facebook user who currently works for the Ministry of Foreign Affairs of Ethiopia, disagrees with Tamiru Addis for instance. He believes Prime Minister Abiy is doing what he can so that the country can reconcile and live in peace. Getachew has no doubt "the young, inspirational leader" Abiy will help Ethiopia tackle internal conflicts once in for all, if citizens follow and support his initiative. The problem, according to him, comes within his own party and party leaders, local administration, who aren't well connected with his ideals. He believes it is time for the Prime Minister to show them who the 'boss' is. Moreover, Nigatu believes the main challenge Abiy faces is to end the dividing ethnic federalism, (a federal system of national government in which the units are separated based on ethnicity) which he says is a must, if Ethiopia shall remain conflict free in the future. Nigatu illuminates his concerns;

"I hope Prime Minister Abiy and his party prefer to go against Ethnic Federalism and set up federalism based on Geography. We are all afraid about the future of Ethiopia and where we are going with the current trend. Ethnic extremists are saying and doing

what they want in the country. Whether we like it or not Ethiopia is very fragile now. The Prime Minister needs to show who the leader of the nation is for these extremist groups and individuals. From city administrations and Woredas, the level of corruption is out of control. Land administrations in Addis Ababa and other major cities around the country are centers with deep corruption levels. Ethnic based corruption is rampant and far from the ears and eyes of the Prime Minister and top officials. Citizens are suffering paying bribes for public servants and individual brokers just to receive basic services. The youth isn't well guided and at the same time unemployed. Extremists are using the youth for their agenda by creating conflicts and instability. If this is not solved as soon as possible and if we fail to funnel the youths' energy and focus to development, the cost would be unbearable. Separating the youth from extremists is very important to bring stability and unity in the country."

Maladministration, corrupt practices, failing government systems as that of the ethnic federalism of Ethiopia; that has instead of giving equalities to all minority and majority ethnic groups and communities in the country as well as freedom and economic well-being, 27 years in practice are pushing way for the proliferation of hatred and violence amongst competing ethnic groups. The highly polarized social media interactions amongst different ethnic groups are evidently exacerbating the violence as observed by my informant Nigatu and others, whereby activists purposely promote hate speech and call for violent actions. Other root- causes of ethnic related violence is further discussed in the next section with proposals of counter initiatives.

The appointments of independent citizens to the National Electoral Board, Supreme Court, Human Rights Commission, sharing the cabinet seats equally with women and his willingness to work with opposition political parties are some of the decisions Abiy took towards democratization. The amendment of the constitution that many Ethiopians believe divides the nation has not happened yet and many hope it will soon be given attention prior to the 2020 elections.

Ethnic tensions and clashes, displacements and ongoing political disputes are still the day to day news in Ethiopia. What is the actual driving force behind growing ethnic extremism in the Ethiopian context? Despite a current internet penetration level of only 15.4%, how has the Ethiopian social media platforms became prominent in shaping political actions on the ground? The next section investigates the root causes and drivers of ethnic extremism and the relationship between actions on the ground and social media activities.

4.3 Causes and drivers of ethnic extremism: The online & offline link

Some of the causes of ethnic extremism in the Ethiopian scenario have been outlined in previous discussions. As I have briefly mentioned in chapter two, the TPLF led EPRDF group that has led the country for 27 years has seen the flourishing of identifying oneself based on one's ethnicity and previous nationalism sentiment of being Ethiopian has been diminishing. The youth in Ethiopia are increasingly paying more attention to associating or affiliating on ethnic lines, creating a polarized view of labeling that has seen the growth of 'us' and 'them' mentality. Such ethnically brainwashed youth with no job and adequate schooling are highly vulnerable to be radicalized on social media platforms. This is mainly because vulnerable youth often look for groups that they can follow who can voice their frustrations in life and easily begin to affiliate to extremist ideals and groups. They feel a sense of belonging and buy into the false narrative of the extremists because they feel society has let them down for not having a job or adequate schooling.

My first informant, Andualem Sysay, says he knows of young members of *Digital Weyane,* a group that is sponsored by the TPLF with the view of advocating digitally to achieve political goals of the party. The group believes the Tigreans' people's culture and values have been under attack by other ethnic groups like the Amhara and Oromo. Moreover, *Digital Weyane* aims to exert political influence digitally by creating its own narrative.

A document obtained from Sysay that contains explanation from the TPLF party as to why there is a need to establish *Digital Weyane* states:

> "Digital Weyane aims to achieve two kinds of political goals. The greatest benefit for Tigray is to effectively protect it from aggression. Like any other peoples of Ethiopia, Tigray has its own political advantage it needs to secure. This ownership, however, is being undermined by those who are working towards political profits. The people of Tigray has never seen attacks at such scale because of their identity in history. Digital Weyane will work towards effectively defending the national interest of the Tigrean people. The digital revolutionary enemies have renewed themselves as part of revolutionary struggle with competing technology and design to influence the youth. As such, it is almost like a development army with a regional mindset and a country. The digital TPLF operates as a structure without being organized. It is a generation that has renewed its struggle form. In addition to protecting the national interests of

> the region, digitalization is a scientific struggle that is actively working to build a developed country with a tolerant, non-conflict society."

According to Sysay, youths deployed by TPLF are mainly fresh university graduates that are paid a trifling amount of money to engage in false propaganda on social media to discredit Abiy Ahmed's government and the reform process. Moreover, the statement from the TPLF points out to unidentified foes which it refers to 'digital revolutionary enemies'. Although the statement does not specifically identify who they are, following up on their discourses that often tends to have ethnically extremist views being reflected on social media, they are addressing mainly the Oromo, Amhara and Ethiopian nationalist groups, who were highly in disfavor of the TPLF and relentlessly unveiled their human rights abuses and corruption, using mainly social media platforms and other radio and television channels like the Ethiopian Satellite Television (ESAT) and the Oromia Media Network (OMN). Sysay said;

> "TPLF's statement is inflammatory in the sense that they have identified those that don't support their parties as enemies. The TPLF has not yet come into terms to acknowledge that it has lost power in the central government. The old TPLF members are sitting in Mekelle (Tigray's regional state capital) exasperated and frightened that their loss of power to the ODP can eventually lead to holding them accountable for past crimes. Although these men do not formally have political power, their previous networks and with the support of Tigray regional government, they are recruiting groups like Digital Weyanes that purposely orchestrate violence amongst groups on social media platforms to destabilize the country and abort Abiy's reforms. Establishing groups such as Digital Weyanes have not only increased extremist voices from other ethnic groups mainly the Amhara, Oromo, Somali, and Sidama; who are engaged in war of words defending the narratives of the TPLF, it has also seen the proliferation of trolls and bots that are used to further disseminate their propaganda and recruit more followers. Sometimes it is very hard to identify the trolls used by TPLF but more often than not they are noticeable and usually make comments under TPLF activists in view of support and heighten the attention posts get by sharing and sending propaganda to different Facebook groups."

A leaked Facebook post shared by Sysay has revealed that the strategy of employing *Digital Weyanes* for propaganda activities is not something new for TPLF. A document obtained from Andualem that was first leaked on social media in November 2017, angering government officials at the time, shows a list of individuals, the exact amount they have been paid and their job titles which the government referred to as 'social media commentators' with the purpose of

promoting the then TPLF led EPRDF. Amongst the list are notable hard core activists of TPLF on social media, Fitsum Berhane and Dawit Kebede, both activists have been accused of actively disseminating a pro-government information campaign during a heightened political tension.

ተ.ቁ	ስም ዝርዝር	ዝሰርሑሉ ትካል	ኣብ ዝቐረበ ምጉጓም	መሃያ	ምርመራ
1.	ዘርኣይ ሃ/ማሪያም	ብሄራዊ መረጻኢታን ድሕንነትን	****	12000	ኣባል
2.	የሞነ ንብረስላሴ	ቢት/ፅ ኢህወዴግ	**	8000	ኣባል
3.	ኣለም ሰገሰ	ትካል ዜና ኣገልግሎት	***	4500	ኣባል
4.	ዜርሁ ሊላይ	ኮርፖሬሽን ብሮድካስቲንግ ኢትዮጵያ	***	7800	ኣባል
5.	ፍፁም ብርሃነ	ቢት/ፅንዳያት ኮሙዩኒኬሽን መንግስቲ	**	8326	ኣባል

መማረፂ ክልተ፡

ኣብዚ ዓውዲ ዝሓሸ ክእለትን ምንቅስቓስን ዘለዎም ግን ድማ ድማሪ ዘይብሉዎም ኣባላትን ደገፍቲን እየ።

ተ.ቁ	ስም ዝርዝር	ዝሰርሑሉ ትካል	ኣብ ዝቐረበ ምጉጓም	መሃያ	ምርመራ
1.	ተ/ሀይማኖትማርግይ	ኮርፖሬሽን ብሮድካስቲንግ ኢትዮጵያ	****	9000	ኣባል/ ማ/ኣመራርሓ
2.	ስናይት መብራህቶ	ብውልቀ	*		ክእለታ ፅቡቕ ዩ ግን ናብ ቢዝነስ ተትኩር እያ።
3.	ጳንኤል ካሳ	ስራሕ ዘይብሉ	*		
4.	ሙሉቀን እማረ ማርግይ	ብውልቀ	*		ኣባል
5.	ባሻ ደሲታ ገ/ሃወርያ	ተማሃራይ (ሕክምና 3ይ ዓመት) ተመሃራይ ስለዝኾነ ምባር ኣብቶም ቀዳማይ ንጅስ ዝምደብ እዩ።	*		ኣባልን ወዲ ተጋሻላይን

Fig. 5 shows the list of 'ten commentators' that were each paid at least 300 USD for blog posts or Facebook messages defending the TPLF led EPRDF coalition at the time. It was first leaked on social media, according to my informant Sysay.

The causes of ethnic extremism in the Ethiopian scenario is obviously not limited to the actions of the TPLF by employing as what the officials regard as 'digital soldiers' as members of *Digital Weyane* that purposely engage on ethnically radicalized narratives on social media. As mentioned before, political moguls are able to politicize longstanding social inequalities amongst groups mainly using social media platforms. One may wonder however, how social media can be so impactful as to cause on such changes in the Ethiopian social, political and economic life despite the fact that only 15% of the population have access to the internet?

My second informant, Endale Assefa explains that in Ethiopia the main problem is many view social media as a legitimate media source and what they would consider as gossip if they heard

it by mouth, when they read about it on social media they take as fact. Journalists, activists, media organizations now release their news either live or breaking using their social media pages, mainly Facebook. This has made Facebook an ideal platform as a source for breaking news stories for many Ethiopian users. Assefa said;

> "Whenever a news breaks that has a potential for many follow up stories, even those that I have talked to in the countryside ask their friends and relatives about what is being said on Facebook even if they don't have access to the Internet and social media. I was very surprised to talk to some young people in the Oromia region that told me they follow up advices and orders from what they consider as respected elders on Facebook and are attentive to posts from activists like Jawar Mohammed for instance. This shows us that the effect of Facebook as a medium and its potential to share videos, photos and documents that can be instrumental as political weapons go beyond social media platforms. Facebook is increasingly becoming a preferred medium for Ethiopian youth especially in towns and cities. Elites from different ethnic groups, the diaspora are usually the ones who are vocal on social media and they are the opinion makers, as they can afford to have uninterrupted internet connections. In addition, the local private newspapers are highly attentive and publish articles taken from social media regularly. Likewise, local FM stations usually pick their agenda based on social media buzz about a certain topic. Starting from the positive aspect of the impact of social media in pushing for the reforms in Ethiopia, on a negative note if extreme violent narratives continue to proliferate on social media and are not handled properly, they will have an immense impact in creating a breeding ground for a great number of violent extremists from all sides. I am concerned both government officials, oppositions as well as the public are growing more intolerant showing the tendency of proposing extreme actions."

Assefa's experience is interesting as it conveys that activities on social media affects offline actors who are not necessarily users of these platforms. These actors are however affected by decisions made by activists and politicians on social media that tend at times to reflect radicalized views. Moreover, it reminds counter insurgency efforts on social media to also consider the impacts of other platforms like the FM radio stations Assefa mentioned in affecting the activities of people on social media platforms.

The results of this study in general have shown that the characteristics of Facebook and its ability to spread information quickly to a range of users have led to violent actions; for example as mentioned in Chapter two, in the case of the medical professionals who were killed by rumors

that went viral on social media that they were in the Amhara region to poison children. This was found to be false news that was too late to for its impacts to be countered.

The power of social media in fueling the polarized politics, lack of know-how about the usage of social media, long standing historical and social injustices, the ethnicity-based federal system of the country, and lack of good governance are all factors that need to be given focus to solve the rising violent extremism in Ethiopia, that is mainly ethnicity-based but at times have turned religious as well. A case in point where ethnic extremism have intersected with religious tensions between Muslims and Christians in the Southern part of Ethiopia, attacks have taken place several times in 2019 in the town of Alaba Kulito, where it was reported that around ten churches were attacked by Muslims after fake news on Facebook had spread about attacks on churches. (Spencer 2019)

In the next section, I try to understand and make sense out of the emergence and growth of ethnic extremist ideologies in the Ethiopian context. Are the ideologies the result of the major flaw of the ruling party's ethnic federalism politics?

4.4 Emergence and growth of ethnic extremist ideologies

The emergence and growth of ethnic extremist ideologies in Ethiopia are strongly linked to ethnic competition entrenched in and promoted by the TPLF-led EPRF government since its establishment in 1991. Ethnic politics has been the trade mark of the TPLF, which has in turn upheld identity based politics affecting all sectors of life in the country. For example, not only was ethnicity a factor in politics but in business affiliation and ownership as well as certain groups like the EFFORT conglomerate owned by the TPLF are run and owned mainly by Tigreans, or TIRET owned by the sister party Amhara Democratic Party (ADP), run mainly by Amharas are to name some for instance. Private Banks and insurances as well as other businesses have also in the past two decades been setup along ethnic lines and affiliations.

To make matters worse, the idea of identity politics is protected in the constitution. For example, the controversial Article 39 (1) to (5) gives full rights to every ethno-cultural community to have its own territory and the right to "a full measure of self-government which includes the

right to establish institutions of government in the territory it inhabits and to equitable representation in state and Federal governments. It also convenes unconditional right to self-determination to ethno-cultural communities - including the right to secession." (Lex, n.d. 12) The constitution defines Ethno-cultural communities as "a group of people who have or share a large measure of culture or similar customs, mutual intelligibility of language, belief in a common or related identities, a common psychological make-up, and who inhabit an identifiable, predominantly contiguous territory." (Lex, n.d., 4)

My fourth informant, Getachew Nigatu of the conviction that:

> "Constitutional amendments are a must. I still can't believe Article 39 is hanging around-this article plus other amendments are much needed and it must happen soon. The constitution has really encouraged ethnic policy to pursue. It has in a way redefined Ethiopia along ethnic lines and has created conflicts of its own especially over the demarcation of territories, competition for resources such as land of the different federal units. There have been several border conflicts over Raya and Welkait communities between the Amhara and Tigray, disputes over border territories between the Oromos and Somalis; deadly conflicts following the people of Sidama zone in the south that have requested to become an independent regional government were at loggerheads with other ethnic group like Walayetas residing in Hawassa, that had at instances turned violent and has claimed many lives. Most of these conflicts have in part the constitution to blame. This is mainly because the federal constitution did not just acknowledge ethno-linguistic identity but created regional states based on those identities. As a friend of mine said the Ethiopian people were not made into the constituent units of the Ethiopian federation but rather ethno-cultural communities. Through school curricula, social programs and several activities the constitution has been promoted for nearly three decades now. This generation of youth have been indoctrinated or brainwashed of the contents of the constitution that identifies territory with ethnicity. This ethnic federalism in general has made people to think of themselves as Amharas, Oromos, Tigres, Guarges, Somali… and not as Ethiopians."

One of the most contentious issues as Nigatu raised in Ethiopian contemporary politics is whether to behold or amend the constitution of 1993 that many commentators believe have been highly influenced by the TPLF led EPRDF. However, not all Ethiopians agree with his view that the amendment of the constitution is a necessary step to guarantee peace in the country. Some Ethno-nationalists that are considered extremists mainly from the Oromo and Tigre ethnic

groups highly oppose the amendment, as the current constitutions gives regions enormous autonomy. For instance, as Nigatu mentioned, the controversial Article 39 gives regions autonomous rights to the extent of declaring their own independent republic country. Although declaring own republic is pragmatically a daunting process, frustrated TPLF hard core supporters, politicians and some advocates of the Oromo Liberation Front (OLF), (a once-outlawed group that entered the country in 2018 following the power shift) strongly oppose any change to the constitution and warn that the country will face civil war if any amendment is made. Extreme activists from these groups often incite fear amongst the public on social media by threatening to call for referendum to have an independent country unless the central government abides by their demands.

My fifth informant **Mebrahatu G. Mariam,** a lecturer at Mekelle University, supports the TPLF and strongly opposes Abiy Ahmed's administration. He strongly believes there's no need whatsoever to amend the Ethiopian constitution.

> "It is a constitution that enshrines the rights of all ethnic groups including minorities. Abiy's government has continuously undermined and violated the constitution. Amending the constitution will not be as easy as Abiy claims it to be. There is no doubt that the TPLF takes the lion share for its sacrifice for the inception and practicality of the constitution and will continue to fight to protect it from being touched. I know Abiy is supporting the voice of the Amhara, who feel they were not represented when the constitution was drafted and adopted. This is a false narrative of the Amhara ethno nationalists that always aspire a unitary form of government that would force people to use one language, one culture and only one way of thinking, just as the previous Military regime of Mengistu Hailemariam and Haile Selassie's imperial government. Not only Tigreans are totally against this but all the other 80 ethnic groups."

Mariam's view reflects the majority of TPLF led EPRDF supporters, who are totally disappointed by the reform which they believe have isolated them. They fear that Prime Minister Abiy aspires to have a unitary state of government, which they claim their party has sacrificed a lot to fight against in the past. TPLF supporters are of the opinion that the Prime Minister's reform will reverse the enshrined equality given to all ethnic groups in the country.

All these political mayhems of the TPLF led EPRDF government in the past decades have led to a polarized nation that is highly divided on ethnic lines. Extremists have exploited ethnic cards for political and economic benefits and mainly thrive on social media by pushing out

rumors and inflammatory claims. As Kebour Ghenna, observed on a Facebook post; “both the theories and the practice of divided identities and dual representation in the Ethiopian federalism have become a key target of nationalists, and especially of Oromo, Amhara, Sidama nationalist elites seeking to monopolize the voice of their people. From their perspective, the ‘Ethiopian’ civic identity of the country as a whole is a threat and a rival.” He added, “Indeed, there are many who describe the federal system as a threat because it divides Oromos or Amharas against themselves.” (Ghenna 2019a)

The emergence and growth of an ideology that proposes ethnic affiliation as the utmost important element in public life has promoted radicalized groups, who endorse and aspire whatever means to achieve their political and economic interests, often resorting to violence themselves and seek to influence others to do the same. As if the alarming tendency of the growing ethnic extremism ideology is not enough, the high unemployment the country faces, an increased inequality amongst groups as well as individuals and political groups that manipulate ethnic differences are all the additional ingredients for violent extremism to prevail. The next section reviews unemployment, social disparity and propaganda as recipes for violent extremism.

4.5 Unemployment, social disparity and propaganda as recipes for violent extremism

The majority of Ethiopia’s 108 million people are below 24 years of age as they make up 63 percent of the population, which is expected to grow to 190 million by 2050 and 250 million by 2100. (Lie 2018, 14) Although the economy of Ethiopia has seen near double-digit growth for nearly a decade, some scholars predict that the growth will not likely keep pace with the swiftly snowballing population. The economy must find employment to thousands of students who graduate from 40 public Universities and numerous colleges every year. In the past years, higher academic institutions were not free from ethnic violence themselves. In fact, several reports have shown that ethnic related violence have steered to several casualties including some students being killed on campuses around the country.

Increased usage of social media in the country means more exposure to radicalized echo chambers, as most Ethiopian users are exposed to extreme groups often representing their ethnicity or a rather radicalized Ethiopian nationalism. There are also growing radicalization on

social media amongst groups that support Ethiopia's ethnic federalism against those calling for a more unitary system of government. As Samantha Bradshaw, a researcher at the Oxford Internet Institute rightly said, "Social media tends to empower propaganda and disinformation in really new ways." (Alba and Satarino 2019)

In a country that is undergoing major reforms after decades of autocracy rule, propaganda targeted at vulnerable unemployed youth expecting immediate benefits from a reform process that strives to attain social equity and justice poses great danger for peace. The youth have in turn become agitated and impatient of the reform they have anticipated would change their lives right away, making them more vulnerable to fall for the narratives of extremist activists and groups. This is mainly because radicalized activists have the tradition of manipulating information to meet their political objectives and use the frustrations of the youth for their advantage by relentlessly trying to convince and inspire them to engage in violence by blaming other ethnic groups and the government for their lack of jobs as well as filling them with propaganda that their ethnic group is under attack.

One of the common tactics used by some extremists on social media is the manipulation of photos and documents such as the one shown in *Fig. 6* with the list of *Digital Weyanes* that are formally paid to disseminate propaganda on social media. Tactics used by extremists such as *Digital Weyanes* range from faking letterheads from a Ministry office from the government and changing the content to fit their narratives to using gruesome pictures from Rwandan genocide or mass killings in Congo and making it look like they have occurred in Ethiopia by writing captions and provoking comments under the pictures. Using such phony images they call on actions from their followers to resort to violent means to defend their ethnic group.

Fig 7. Getachew Shiferaw, a prominent activist for the Amhara ethnic group has exposed extremist Ethno-Oromo activists that have circulated a picture from the Rwandan Genocide and used it to escalate the violence by claiming the victims shown in the picture are from the Gumuz ethnic group attacked by the Amharas.

My informant Tamiru Addis attests:

> "I have witnessed several times that some trivial conflicts amongst individuals have been made to appear as ethnic conflicts on Facebook and people have falsely been driven to violence at a larger scale because of the propaganda. Activists that use photo shopped pictures to deceive followers are many in number. For example, Ethno-Oromo extremist activists have used a picture from the genocide of Rwanda (see Fig.7) claiming that they are victims of the Gumuz ethnic group that were killed by the Amhara ethnic group in Benishangul Gumuz state. It's true there were clashes between the Gumuz and Amhara ethnic groups but the extent of the clash has been exaggerated by the Oromos to blame the Amhara for the killings as well as to exacerbate the violence. Similarly stories covered by social media pages such as Jijiga Herald used an image from a conflict that happened a couple of years ago in Burundi and made it appear as if it is an orchestrated massacres of Somalis in Oromia. The photo shows a truck carrying several dead bodies in Burundi with the aim of exaggerating Somalis victimhood and inciting

them to retaliate against the Oromo ethnic group. I am not saying that there were no conflicts and casualties between the Somalis and Oromos at the time, but it is saddening that some activists were committed to adding fuel to the violence by using images taken from other conflicts."

Addis's testimony about the deception tactic used by violent extremists to invoke terror is a reminder that such misleading photos and videos used purposely by extreme groups and individuals can have ultimate impacts in exacerbating exciting conflicts or inflicting fresh ones. I will in Chapter Five explain how such false news on social media can be detected and prevented by implementing AI powered counterinsurgency models.

On the other hand, some observers like my informant, Getachew Nigatu, are of the opinion that the government of Abiy Ahmed has not yet placed a tangible strategy to counter the proliferation of false news and propaganda. He says a strategy design to combat extreme violent narratives on social media is essential. But Nigatu is against the current draft law recently proposed by the government targeted at placing hefty fines on social media users that engage into hate speech and terror on these platforms. Nigatu said;

"I am actually against the new draft law which will certainly jeopardize the freedom of speech. We have seen it in the past. The EPRDF has a bad reputation when it comes to manipulating such laws, I don't believe that tradition will go away real soon. The current criminal code of the country clearly penalizes those that disseminate false information, aggravate hate and violence. We need to focus instead to empower the youth by investing more on programs that would enroll the majority to jobs and intensify community programs and work on creating a culture that is tolerant to accept differences. The *Medemer* (synergy) philosophy of Prime Minister Abiy can be used to teach the youth about reconciliation, unity and walking together as a nation. An informed society will not need to be policed by the government just for exercising their freedom of speech on social media. The youth must know, no one is above the law. What's happening every day is alarming as some groups are intentionally working to weaken the government of Ethiopia. Some people are losing trust in the government's willingness or ability to control the situation. We should not forget that Al-Shahab, Interahamwe, Janjaweed Militias or ISIS were formed by clueless, uneducated, unemployed youth mass. We don't want to see the formation of such terrorist groups amongst our youth."

Unemployment, social disparity and propaganda are certainly recipes for the flourishing of chaotic violence in the country. The named factors as I have tried to signal above have the ability to directly or indirectly aid the flourishing of radicalized voices that may potentially lead to vehemence and conflicts.

Moreover, the reform is faced by many obstacles including the breakup of the ruling party EPRDF coalition that has long been dominated by TPLF. Now that Abiy's party ODP controls the coalition and the central government in Addis Ababa, angry TPLF officials have left Addis, and spend considerable time in the town of Mekelle, Tigray regional capital, charting potential responses to the reform process. TPLF officials have made several official statements contradicting and opposing Abiy's government that have often the tone of anger and frustration with the central government. Some also fear the feud between TPLF and the central government may get out of hand pushing the country into a civil war. Others have a more optimistic view and believe dialogue and negotiations will solve the political deadlock Ethiopia is in at the moment. One of my informants Endale Assefa said;

> "I support Abiy because I don't see any other politician of his caliber who will maintain a united and proud Ethiopia that can guarantee equal rights and freedom for all groups. We are sick and tired of parties and structures that only favor a certain group along ethnic lines and affiliations. It is now time for all of us (Ethiopians) to value virtues based on content of character instead of ethnic background. If central political and economic disparities can be addressed, the state will not be the one that punishes those that disgorge hate and propaganda. It would be society, made up of an informed public that rejects and pushes anonymity, misinformation and hate speech."

My informants in general paint a picture of the contemporary Ethiopian social network atmosphere, where 27 years of ethnic oriented political standoffs have become more exposed and heightened between Ethno- nationalists groups. This has also included constituent parties within the ruling coalition, EPRDF, divided on ethnic lines, at times followers of the party reflecting radicalized views on social media. Especially activists that tend to exhibit extreme behaviors mainly from the Amhara, Oromo and Tigray ethnic groups have dominated activities on social media platforms mainly on Facebook.

The opening up of the political space after decades of repression has led manipulative social media activists to take advantage of distorted historic narratives to create more polarization amongst groups using several tactics like disseminating false news. The Ethiopian reform also

faces challenges on sensitive issues like amendment of the constitution that have divided radicalized Ethno-nationalists against some Ethiopian nationalists. With a significant number of restless unemployed youth, Prime Minister Abiy seem to be cornered from all angles ahead of the 20 20 elections. Amhara extremists accuse him for advancing the interests of the Oromos first and not taking actions against extreme Oromos because he is an Oromo himself. The Oromo extremists accuse him of not advancing their interests and display hate to his Ethiopian rhetoric. Extreme Tigreans and angry TPLF officials unhappy about losing power in central government have accused him of isolating them and taking actions against their interests. The TPLF has sponsored groups like *Digital Weyanes* to abort the reform led by Prime Minister Abiy Ahmed.

It is most of all interesting that the study reveals activities on social media affect offline actors who are not necessarily users of these platforms. These actors are however affected by decisions made by activists and politicians on social media that tend at times directly impact actions on the ground as I have tried to signal through the previous examples. This perhaps explains why countries with such poor internet infrastructures like Ethiopia, social media still plays a huge role, as it is a popular medium amongst the youth and communities. In addition, social media platforms are highly used by local newspapers and FM radio stations as major source of news and social media buzzes are influential to set agendas for traditional media houses in Ethiopia.

In the next chapter, conclusions of this project are discussed in light of relevant theoretical frameworks outlined in chapter one. I have also summarized my findings and share prospects of further researches that are related to non-human actors that can be effective as identifiable counter insurgency models in battling social media radicalization that may lead to violence.

Chapter Five

5. Conclusion, discussions, summary of findings, further researches

5.1 Identifying violent extremist individuals and groups on social media

In a global society of 7.7 billion people (Raftery, Alkema, and Gerland 2014) ingrained with injustice, greed, delusions, and political upheaval, various anomalous extremists are bound to exist. It is now evident that violent extremism has gained unforeseen momentum with the advent of the social media. As exemplified in Chapter Four, it is no news now that the social media has been added to the paraphernalia of tools used by violent extremists to further their domain. For instance, violent ethnic extremists as I have discussed in the case study of Ethiopia, exploit the social media platforms convenience for free speech to radicalize individuals by instigating hate-mongering and glamorizing violence.

However, identifying violent extremists is not always an easy task. In-depth single case studies like that of Ethiopia's social media driven collective actions are valuable for researchers to delve into complex systems in which outcomes are determined by the interaction of numerous factors and thus offering plausible casual explanations. As (Schmid 2013, 35) rightly said, "radicalism and extremism can be properly assessed in relation to what is mainstream political thinking." Schmid's theory of radicalization that understands the concept as "a context-bound phenomenal, where political, sociological and global matter as much as ideological and psychological ones" is important to incorporate in efforts related to ascertaining violent extremist individuals or groups on social media.

My informants' understanding of radicalization for instance differs considerably in the sense that some of them accuse each other of being an 'extremist' pushing for and inspiring violent actions. Moreover, social media companies like Facebook that reinforce take-down and ban policies against extremist material provoking violence are either too slow to remove them or just don't notice the material. Even thousands of pseudo accounts, bots and evidently dangerous news pages continue to exist in Ethiopian social media networks without any action taken

against them by social media companies. According to my informants, extremists use several tactics on social media cleverly so that they are not spotted easily and labeling them as potentially dangerous can be at times difficult, even though they often resort to romanticizing violence in devious ways that would invoke young users to engage in the process of radicalization and eventually aspire them to commit extreme violent actions.

In addition, social media platforms are conveniently designed to create multiple opportunities to channel violent extremism content reaching thousands if not millions of followers simultaneously without attracting that much attention from content reviewers of social media companies. The free and uncontrolled nature of these platforms enable radicalized extremists to form groups and disseminate their beliefs. Some features on Facebook, for instance, the ability to send live videos to millions of users at once would practically make it difficult to follow and disrupt all contents disseminated by violent extremists unless tipped by concerned followers.

A case in point where a violent extremist has used Facebook without being noticed is what happened on March 15, 2019, Brenton Tarrant, a 28-year-old Australian, who describes himself as a 'fascist and white supremacist', reportedly published a racist manifesto before he opened fire at two Mosques in Christchurch, New Zealand killing 51 people in a Facebook live-streamed video attack.

Similarly, inspired by the New Zealand attacks, a gunman opened fire at a crowd in early August 2019 at a shopping complex in El Paso, Texas, killing 20 people and injuring 26. The suspect, Patrick Crusius, a 21-year-old white man from Allen, Texas, draws direct inspiration from the New Zealand mass murder and before he allegedly opened fire, a hate-filled, anti-immigrant manifesto appeared online that spoke of a 'Hispanic invasion of Texas' detailing a plan to separate America into territories by race warning that white people were being replaced by foreigners. The manifesto that may be linked to the suspect, if proven by police, shows the concerning trend of social media effectually being used by growing white supremacists to promote violent extremism and terrorism globally without being stopped by social media corporations (Arango, Bogel-Burroughs, and Benner 2019). It is important to note here that social media corporations are equally important actors in any conflict that may arise because of their decision-making abilities to prevent such violent narratives and their action or inaction can play a huge role in conflicts.

The Counter Extremism Project (CEP) at a recent report (Andrea Page 2019) asserted that Facebook is failing to enforce its own Community Standards for allowable content by failing to remove pages representing extremist groups. The CEP contends in the report that "Facebook only monitors sites in reaction to complaints of extremism and hate speech, does not have an adequate site reporting system, and is not proactive enough in preventing extremist group from operating sites on its platform." The report further calls on Facebook to do more to stop radicalization in social media. "Facebook must improve its reporting features and analysts training so that questionable Violating Community Standards are identified and removed more readily." Stated the report. (Andrea Page 2019)

Understanding and predicting human behavior using information from people's activities on social media can in theory be possible. It is however inconceivable if we were to manually analyze several hundreds or thousands of users activities without the help of machine analysis, as one user may generate thousands of words in a post and each sentence can be respectively multifaceted.

As I have mentioned in Chapter One, non-profit organizations such as the Institute for Strategic Dialogue (ISD) have taken advantage of the progress in Artificial Intelligence, particularly the revolution in machine learning to pinpoint violent extremists on social media. Natural Language Processing (NLP) is such a system that no longer requires the simple interpretation of a text or speech based on its key words but instead have the ability to 'cognize'. NLP is a system that is able to understand meaning behind words, including detecting figures of speech like irony.

Diego Lopez Yse, in "Your Guide to Natural Language Processing (NLP), How machines process and understand human language*"* defines Natural Language Processing (NLP) as a "field of AI that gives machines the ability to read, understand and derive meaning from human languages." (Yse 2019, 5)

Yse explains that verbal and written expressions carry huge amounts of information that can be construed and value extracted from to predict human behavior on social media. NLP amongst its several applications in other fields can also be used to identify fake news.

An alliance of a team of computer scientists at MIT, CSAIL and QCRI, (part of Hamad Bin Khalifa University), for example, have developed a system using data from Media Bias/Fact Check, (MBFC) a website with human fact-checkers who analyze the accuracy and biases of more than 2,000 news sites; from MSNBC and Fox News; and from low-traffic content farms.

They then fed those data to a machine learning algorithm with the intention to classify news sites the same way as MBFC. (Conner-Simons 2018)

The researchers found out that "when given a new news outlet, the system was then 65 percent accurate at detecting whether it has a high, low or medium level of factuality, and roughly 70 percent accurate at detecting if it is left-leaning, right-leaning, or moderate." The team observed that the most unfailing ways to detect both fake news and biased reporting were "to look at the common linguistic features across the source's stories, including sentiment, complexity, and structure." (Conner-Simons 2018)

Although such technologies that use AI to combat the spread of false news have not yet fully been developed, they however signal promising results that can potentially in the future be deployed to haunt fake news disseminated by violent extremism individuals and groups in the case study of Ethiopia, for instance.

My fourth informant Getachew Nigatu says the use of advanced technologies such as the Natural Language Processing (NLP) to identify individuals and groups that work relentlessly to attract young people to be radicalized by continuously aspiring violence must be challenged and removed from social media platforms. Nigatu advices;

> "The government should constitute AI as a central means to quash ethnic tensions in the country. This is mainly because groups who are against Prime Minister Abiy's reform, like the leadership in Tigray, the TPLF, upset that they are pushed out of the central government, are doing and will continue to cause mayhem intentionally by investing on technology to push propaganda. They have evidently deployed bots and trolls as well as organizing what TPLF calls 'digital soldiers' such as notable groups like Digital Weyane that promote violence to abort the reform. We should also remember that the TPLF has the resources and means to destabilize the country and it is not the only group. Extremist factions in Oromo Liberation Front (OLF) as well as National Movement of Amhara (NAMA) have also demonstrated alarming violent activities on social media. Ethno extremists from all sides must be clearly identified first and their bigotry behavior must end by using AI and other counter strategies."

As Nigatu observed, such deployment of bots and trolls by the TPLF further complicates the process of identifying violent extremists. One of the potential problems that are associated with such malicious bots is their ability to create false positives, further complicating the process of identifying violent extremists on social networks. Since it is difficult to differentiate these bots

from human users, they can easily penetrate networks and are greatly affecting discourses on social media and are often pushing for radicalized views, aspiring violence on these platforms. Ultra-modern AI-based bot detection system can only keep up with these malicious bots in counter insurgency designs against violent extremism.

Moreover, identifying violent extremists on social media is only the first step to battle the narratives of extremists and push back against recruiting new members on social media. Counter insurgency efforts also need to incorporate a well-thought out counter narrative strategy targeted at individuals and groups who are at risk of engaging with violent extremists.

There have been well established efforts of creating counter-narratives by governments and non-governmental organizations like ISD. However, measuring the effectiveness of such attempts in changing behavior has been a daunting task. In addition, the role of AI such as the deployment of chatbots that can be used for counter narratives cannot be underestimated; as it is in the nature of AI to perpetuate both evil and good. Chatbots can well be used to support the narratives of violent extremists but have an evenly matched ability to counter as well. The next section looks into the role of counter narratives on social media to encounter violent extremism.

5.2 Role of counter narratives to challenge violent extremism

One of the essential aspects of countering violent extremism is challenging the narratives of individuals and groups that tend to push for information that aspires the engagement of violence to achieve political, religious or economic goals. According to my informants in this study, the voices of extremists have dominated Ethiopian social media networks and they thus suffer from a lack of a nuanced perspective that could convince and engage youth to dialogues instead of conflict. The ISD defines a counter narrative as a "message that offers a positive alternative to extremist propaganda, or alternatively aims to deconstruct or delegitimize extremist narratives." (Davey, Birdwell, and Skellett 2018, 4)

A recent report published by Samantha Bradshaw and Philip N. Howard, *The Global Disinformation Order 2019 Global Inventory of Organized Social Media Manipulation*, Computational propaganda Research project, indicates that "Computational propaganda- the use of algorithms, automation, and big data to shape public life- is becoming pervasive and ubiquitous part of everyday life." The report adds, "around the world governments are using

social media to manufacture consensus, automate suppression, and undermine trust in the liberal international order." (Ma 2019, 15)

Starting from 2017, the team has over the past two and a half years monitored the global organization of social media manipulation by governments and political parties. The 2019 report analyses the trends of computational propaganda and how tools, capacities, strategies have evolved to exert influence on social media platforms.

The report presents evidence of organized social media manipulation campaigns which have taken place in 70 countries, up from 28 countries in 2017. "In each country, there is at least one political party or government agency using social media to shape public attitudes domestically." (Ma 2019, 45) In addition, the report reveals how social media has become co-opted by many authoritarian regimes like the TPLF-led ERDF government of Ethiopia and is being used as a tool of information control in three ways, namely, "to suppress fundamental human rights, discredit political opponents, and drown out dissenting opinions." The report stresses how what it calls 'cyber troops' to denote those appointed by governments to exert influence on social media and other cyber platforms work together with many other actors that we need to pay attention to. It states; "One important feature of the organization of manipulation campaigns is that cyber troops often work in conjunction with private industry, civil society organizations, internet subcultures, youth groups, hacker collectives, fringe movements, social media influencers, and volunteers who ideologically support their cause." (Ma 2019, 48)

In the Ethiopian situation, I was able to track three types of fake accounts: bot, human and cyborg. Bots are highly programmed accounts deigned to simulate human behavior online. In general however human-run accounts that don't make use of automation are much more common than bots or cyborgs.

My first informant Andualem Sysay started answering my question on how strong counter narratives are in Ethiopia to combat ethnic extremism with the famous quote from the renowned physicist Albert Einstein that states "the world is a dangerous place, not because of those who do evil but because of those who look on and do nothing."

According to Sysay, the passiveness and silence of many Ethiopian scholars, whose participation is very limited on social media networks have given green lights to violent extremist groups to dominate and control the social media platforms. Sysay suggests;

> "If a systematically powerful counter narrative strategy against those Ethno extremist groups is put in place on social media, they would not only eventually loose followers, the entire culture of Ethiopian social media network that suffers from reflecting highly radicalized views on ethnic lines would slowly tone down. This would gradually encourage users to debate on issues rather than giving so much focus on ethnic identities. They can also learn to adopt other means for solving conflicts other than violence. We cannot afford to keep quiet and see the country disintegrate anymore. Most of the so-called activists are only after gaining fame, money and power at whatever cost, even if it means destabilizing the lives of thousands if not millions of people. We have to let young people that are mainly the targets of violent extremists know that they are being used by these manipulators and the only way we can do this is by battling their ridiculous ideology of Ethno extremism and the crumpling ethnic based federalism of the government. The regional government of Tigray led by TPLF is aspiring for violence themselves; they need to sit down and reconcile both with the Amhara regional government and central government before things go out of hand and civil war eventually breaks out."

There are however, according to my informants, some organized attempts by few individuals and groups to battle the threat of radicalized extremist groups. One such endeavor is carried out by a group called TIKVAH-Ethiopia, formed by some concerned fresh university graduates with the objective of combatting the proliferation of hate speech on social media and expose false news by verifying reports.

The group mainly uses a Telegram page to reach out to their more than 460,000 members. Even though they use other platforms like Facebook, their influence has been limited. Such efforts are appreciated to engage alternatives for youngsters in peace, dialogue and debate, but they are far from adequate to weaken and confront the narratives of extremists. TIKVAH-Ethiopia's main channel Telegram as opposed to other platforms like Facebook, mainly offers the option of interacting with users in a sort of 'filter bubble' in the sense that their followers are willingly in the group and often let like-minded people flock together and exposing them to echo chambers as opposed to sharing and understanding others' points of views. This in a way isolates users in their own ethnic, religious or cultural bubbles having its own implications on free speech and democracy.

Internet activist, Eli Pariser, coined the term 'filter bubble' in his book "The Filter Bubble: What the Internet Is Hiding from You" (2011). Techopedia defines 'filter bubble' as "the

intellectual isolation that can occur when websites make use of algorithms to selectively assume a user would want to see, and then give information to the user according to the assumption." (Techopedia n.d.)

These assumptions made by for instance, social media companies like Facebook that show you updates from friends you interact with the most and filter out people with whom you have less common views promote the proliferation of echo chambers on social networks. According to Pariser, such assumptions based on the information related to the user, such as former click behavior, browsing history, search history and location, will result in a filter bubble which he defines as "a unique universe of information for each of us" (Pariser 2011, 7)

Pariser's concern relates to the lack of random or in his term 'serendipitous' communications on the internet that would instead feed users with information they agree with and are less likely to challenge existing views. "A world constructed from the familiar is a world in which there's nothing to learn," Pariser calls this "invisible auto propaganda, indoctrinating us with our own ideas." (Pariser 2011, 11)

The issue of 'filter bubble' is an important aspect to consider in counter insurgencies against radicalized individuals and groups on social media. Different ethno-radicalized groups in the case study of Ethiopia for example are in their own bubble and often two versions of the same news story narrated by different groups create confusion on social networks.

Often the majority of fake news situations are not easily countered by facts. This applies not only to ethno-extremists in the case study, it can equally apply to other communities like radicalized supporters of the U.S President Donald Trump on social media. Extremist supporters are not convinced by all the information provided to counter fake news; they would instead actively seek out counterfactual material that supports their view. Counter insurgency efforts must bear in mind to burst such filter bubble that surrounds each user especially individuals and groups who voice radicalized opinions and materials on social media platforms. Radicalized individuals, groups that aspire for violence need to be given especial attention due to their potential to recruit youngsters engage in violent extremism. Pariser suggests that internet companies must give users more control over their privacy in the sense that they must let users know the personal information these companies have in store about them.

Pariser's argument of the impact of 'filter bubble' in having adverse effects for social discourse is not universal though. Others say the impact is negligible. An article from *The Economist* magazine entitled 'Invisible sieve' for instance challenges Pariser's suggestion that filter

algorithms could be complemented by human editors who show you worthy things you ought to see will put internet companies like news providers to accusations of bias. The article finds strange that Pariser calls for an "active promotion of public issues and cultivation of citizenship by big internet firms." The article stated. (The Economist 2011)

There is however a very important point worthy of noting forwarded by Pariser that challenging social media companies to change their operations in terms of allowing algorithms control what you will want to see. This practice not only invades the privacy of users but can also be exploited by radical violent extremists to channel messages of violence, to plot and orchestrate terror without being noticed by content reviewers of social media companies.

In counter initiative efforts against violent extremism as I have mentioned earlier in the example of TIKVAH-Ethiopia must give focus to the adverse effects 'filter bubble' has in social discourses. 'Filter bubble' is an important and easily overlooked aspect of the social networks evolution that affects everyone who uses it, especially vulnerable youth that may fall for extreme violent echo chambers.

Aklesia Sisay, one of the managers of the TIKVAH-Ethiopia says they have established the page because of the concerning bigotry, hate and false news on social media that has really impacted her and peers when they were still students at the Addis Ababa University. "Our strategy is not to directly counter the rhetoric of violent extremists but attracting more people to our group by devising creative activities such as quizzes, games and entertainment on social media at the same time empowering the youth to be vigilant against radicalization that would manifest itself to terror." Sisay says.

Such manual efforts to counter extremist narratives could be more effective if they can be able to adopt Artificial Intelligence that is able to generate for instance content on its own and engage in counter narrative efforts. Advances in AI and Machine Learning (ML) has improved the expediency and proficiencies of chatbots and may potentially grow into a powerful tool in battling the narratives of violent extremists. Chatbots are envisioned to become more human-like by learning from human behavior and have the potential not to be constrained by time and space that would make them ideal in counterinsurgency efforts along with human actors.

As N. Katherine Hayles and other scholars have pointed out old-fashioned ethical inquiries that have accentuated on the individual human considered as a subject processing free will are inadequate to deal with technical devices such as chatbots that operate autonomously, in addition to complex human-technical networks in which "cognition and decision-making

powers are distributed through the system," which Hayles refers to the latter as 'cognitive assemblages'. (Hayles 2017, 4)

Through the lens of Hayle's concept of human-technical cognitive assemblages, counter insurgency efforts that have taken into consideration the potential of AI to cognize and their capabilities to learn and adopt from their environments, like the Ethiopian social media network, not only makes these non-actors relevant but essential to defeat the narratives of violent extremists. This is mainly because as it is in the nature of AI that chatbots can also potentially be designed to fabricate propaganda, recruitment and engage the youth to be more radicalized and violent extremists use chatbots to perpetuate fear amongst the public and incite conflict.

(Abdul Rahman and Suguna 2017) for example speak of chatbots that can potentially be programmed with a basic database of extremist narratives and responses to manipulate the psychological vulnerabilities and social grievances of people who are seeking answers and support on social media platforms.

They warn that chatbots in the wrong hands can facilitate security threats. Chatbots can function as robot extremists spreading radical propaganda, supporting or even supplanting human extremists. They may also be instrumental in extremist enterprises to build affinity with their respective target audiences for the purposes of victimization and online radicalisation. Indeed, the ploy of using chatbots that mimic humans for propagating radical views and reaping support is hardly extraordinary given the reported use of pro-Trump twitter bots during the 2016 United States presidential campaign.

Groups such as *Digital Weyanes* in the case study for example, that have the economic and technological means and a blessing from the TPLF led administration of Tigray region have evidently used malicious chatbots. Members have also been sponsored to receive digital training in China to purposely disseminate propaganda that often resorts to ethnic competitiveness as well as radicalized notions of identity that often encourages violence, according to my informants.

They have used the tactic of 'scapegoating' against Prime Minister Abiy Ahmed and other officials of the government, where the *Digital Weyanes* have a tradition to single out particularly the Prime Minister on social media consistently giving him unmerited blame, character assassination and consequent negative treatment for every incident in the country. Even as I was on the final stages of writing this paper, news broke out that Prime Minister Abiy

Ahmed has won the Noble Peace Prize of 2019 for his peace efforts in Ethiopia and the East African region. However, even such news were welcomed with criticism from the *Digital Weyanes* that went on to demean the Noble Peace Prize award circulating false news that the award is a western instrument to anoint Abiy Ahmed to carry on with the west's 'neoliberalism' agenda trying to discredit the award, circulating false news on social media that the likes of Adolf Hitler and Joseph Stalin were both Noble Peace prize winners.

Such destructive forces that aspire to violent extremism to solve political differences require an equivalent counter narrative against ethnic oriented extremism by deploying chatbots that can hypothetically be programmed to generate narratives that emphasizes on engagement in dialogues to solve conflicts, tone down ethnic competitiveness by aspiring for equality and enforce a change in paradigm when it comes to tolerating and respecting others' political views, religion and culture. False news that is intentionally orchestrated by violent extremists to fuel existing violence or inciting a new one as I mentioned in the previous section can be prevented by adopting Natural Language Processing (NLP) that can determine if a source is accurate or politically biased.

A system that makes use of NLP for instance have been studied by researchers at MIT CSAIL, where two different approaches have been applied for integrating AI for detection of false news. The first research focuses on identifying text that has been generated by a machine, as this is often a source of fabricated information and a common vehicle for its rapid proliferation on social media. This approach focuses on the origin of news sources in determining whether the source is malicious.

The researchers have found out a fundamental problem with what they call "stylometry-based provenance" approach, a system that traces writing style back to its producing source and determining whether the source is malicious. This method, they observed, was only shown to be highly effective under the assumption that fake text is produced by a language model. However, in their studies they have observed that fake and legitimate texts can originate from identical sources. The researchers noted that their findings highlight the importance of assessing the authenticity of the text rather than solely relying on its style or source. (Schuster, Schuster, et al. 2019)

The second approach to detect fake news, the research team at MIT CSAIL analyzed fact-verification methods that rely on FEVER dataset (the largest dataset for Fact Extraction and

Verification in text). They have used a different approach to verify claims and show how bias can impede verification algorithms.

"Fact verification requires validating a claim in the context of evidence. We show, however, that in the popular FEVER dataset this might not necessarily be the case. Claim-only classifiers perform competitively with top evidence-aware models." (Schuster, Shah, et al. 2019) The researchers observed.

The researchers investigated the cause of this occurrence, identifying strong cues for predicting labels solely based on the claim, without considering any evidence. They created an evaluation set that avoids these eccentricities. The performance of FEVER-trained models significantly drops when evaluated on this test set. As a remedy, they have introduced a regularization method which "alleviates the effect of bias in the training data, obtaining improvements on the newly created test set." (Schuster, Shah, et al. 2019) Such efforts as I have exemplified are works in progress but certainly considerable steps towards a more sound evaluation of reasoning capabilities in fact verification models.

There have also been other attempts to use NLP to detect the authenticity of news stories on social media platforms and the web in general. For example, (Conroy et al. 2015) have provided an assessment method that uses a combination of 'linguistic cue approaches' and 'network analysis' to assess and detect news. This innovative approach has given promising results. Other researchers such as (Rubin et al. 2015) have also separated the detection of fake news in three categories, namely, large scale hoaxes, serious fabrications and humorous fakes. Their approaches involved analyzing word patterns and statistical correlations of news articles. (Shao et al. 2016) introduced Hoaxy, a platform for the collection, detection and analysis of online distortion, which had the potential to help people understand the subtleties of real and fake news sharing.

However, such endeavors that involve the use of NLP to detect false news since 2015 are still under development and yet to achieve a full-fledged detection method. Moreover, such detection methods have been criticized for focusing on the linguistic aspects of an article and not so much on fact checking. For example, (Zhou et al. 2018) argue that fact checking should be adopted in conjunction with linguistic characteristics analysis to truly separate fake news from real news. They were able to show through experiments on *Fakebook*, a fake news detector considered highly operative, that fact tampering attacks can be effective and can potentially risk

to fall in the hands of violent extremists and propose "a crowdsourced knowledge graph as a straw man solution to collecting timely facts about news events." (Zhou et al. 2018)

With rapid advances in chatbot technology, it will soon probably be not a matter of choice to adopt social bots in efforts to combat violent extremism on social media. As I have signaled in the examples of *Digital Weyane* that consider themselves as 'social media soldiers' that can take advantage of chatbots that can also potentially be programmed with a basic database of extremist narratives and influence the mental vulnerabilities and social grievances of the youth by meddling in real news or other technological means to enrich their control on social media platforms.

Efforts to combat violent extremism on social media must have a basic understanding of how humans and technical systems are interconnected and affect the cognitive decisions of each other in the 'network' as used by some scholars like (Latour 2007) or to use (Hayles 2017) term in the 'assemblage.' The implications of the increasing autonomous nature of chatbots and AI need to be analyzed in order to understand more clearly as Hayles and others have attempted, how their "cognitive abilities interact with and interpenetrate human complex systems"? (Hayles 2017, 8)

In the next section, I look into the integration of Artificial Intelligence in 'moral actions' on social media platforms in an attempt to understand better what this means for counterinsurgency efforts designed against radicalization and violent extremism.

5.3 Incorporating Artificial Intelligence (AI) in 'moral actions' on social media platforms

As I have indicated in chapter one under the section, "Moral actions in digital platforms" the issue of morality must be seen in a new light in order to encompass AI and their dynamic nature of evolving, making them more and more independent agents that have both the potential to make moral or immoral decisions. For the sake of simplicity and relevance for this research, I have borrowed the definition of morality from (van Berkel et al. 2019) from their study on "Contextual Morality for Human-Centered Machine learning" which suggests that AI-powered decision making must align with the core principles of human morality, where they define

morality as “the codes of conduct put forward by an individual, group or society that can distinguish right and wrong behavior and decision-making.” (van Berkel et al. 2019)

By using the four units of analysis that I have mentioned in chapter one, namely moral sensitivity, moral judgment, moral motivation and moral action; processes that function together in fostering the completion of a moral action; for instance; a chatbot AI function that can be hypothetically designed to be informed about these units and learn the causes of conflicts between polarized groups on social media can potentially be adopted to eventually be able to resolve conflict as mediators, moderators or perhaps act as social media moral police.

In a highly ethnically volatile environment like that of the Ethiopian social networks, tracing all the responsible actors and their roles in conflicts is important. In the case study of Ethiopia, for example, Tamiru Addis, one of my informants has eloquently summarized his opinion of the current situation in Ethiopia that he believes are responsible for the polarized political environment that has led to an increase in violent extremist actions based on differences on ethnic lines.

“We have a growing Oromo-nationalism within the ruling party ODP that the Amhara-nationalism became reactionary to, whereas Tigrean-nationalism is fanning conflicts as well as activists seeking justice in mob campaign making Addis Ababa and other cities battlefields, while the federal government has become a toothless dog.” Addis said.

The prospect of AI is fascinating as it can now in theory potentially be developed to cognize such statements and be able to understand them as opinions rather than facts. AI can learn and refer to historical grounds of disputes, engage in facilitating arguments and carry out fact checking. Moreover, AI can be designed for instance to identify central elements that are sources of conflicts amongst views from ethno-extremist groups that aspire and push for violence and take appropriate actions based on its own evaluations; either remove them from platforms or disengage them from the process of radicalization by intervention.

Clever designs can also be informed that can for instance hypothetically are able to run algorithms that take into consideration and analyze what actions are possible, who and what might be affected by each possible action and how the involved parties react to possible outcomes on social media. AI can then further be designed to be able to reason out about the possible actions and deciding which is most moral or ethical and prioritizing what is considered to be the most moral or ethical action over others.

In each stage, AI must be able to be intent upon following the course and finally combining the strength of will, the social and internal skills necessary to carry out the intended course of action. If autonomous agents such as chatbots are designed to accomplish these processes, they would be involved in taking moral actions as the progressions explained function together to develop a moral action as some scholars suggest.

However, carrying out such AI projects is easy said than done. Developing and deploying such AI functions requires enormous scrutiny so that it follows moral guidelines that are ethical, safe and responsible. The prospect that advancement in AI will help humanity to defy some of its most urgent challenges like violent extremism on social media is exciting, but legitimate concerns still abound. AI falling in the wrong hands and manipulated by for instance violent extremist groups and failure in some AI designs that may act the opposite of what they were originally intended for are perhaps some of the major problems that we need to pay attention to in the development of AI for counter insurgency efforts against violent extremism on social networks.

David Leslie, in a manual designed at the Alan Turing Institute under the title '*Understanding artificial intelligence ethic and safety'* for instance warns that "as with any new rapidly evolving technology, a steep learning curve means that mistakes and miscalculations will be made and that both unanticipated and harmful impacts will inevitably occur, AI is no exception." (Leslie 2019, 3)

Leslie has suggested AI ethics guide and has provided the conceptual resources and hands-on tools that will enable designers to oversee the responsible design and implementation of AI projects. He defines AI ethics, "as a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies." Leslie (2019, 4)

Other authors such as in the public document created by committees of the IEEE Global Initiative on Ethics of Autonomous and Intelligent systems, (a global initiative composed of several hundred participants from six continents, who are thought leaders from academia, industry, civil society, policy and government) also contend that "AI should be designed in a way that both respects and fulfills human rights, freedoms, human dignity, and cultural diversity and that AI must be verifiably safe and secure throughout their operational lifetime," ("IEEE SA - The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems" n.d. 118)

The committees recommend legal frameworks embraced internationally for securing human rights that can be "emulated, adapted, and proliferated, regarding ethical best practices around AI to best honor human rights" such as, The Universal Declaration of Human Rights, 1947, The International Convention on the Elimination of All forms of Racial Discrimination, 1965, The UN Guiding Principles on Business and Human Rights, 2011 etc. Designers, according to the document, must "find ways to translate existing and forthcoming legal obligations into informed policy and technical considerations is needed. Such a method should allow for differing cultural norms as well as legal and regulatory frameworks." ("IEEE SA" n.d., 112)

Incorporating legal frameworks in AI is important and may perhaps be the first step to counterinsurgency efforts against violent extremism on social media networks. Laws would not be difficult to identify as they are public documents. On the contrary, the adaptation of AI in moral actions are much more challenging to determine as they are expressed through language, customs, artifacts, cultural symbols etc. and differ accordingly in various communities. The committee through their studies on Ethically Aligned Design argue that "generating a universal set of norms that applies to all autonomous systems is not realistic, but neither is it advisable to completely personalize an AI to individual preferences." ("IEEE SA" n.d., 132)

Even though autonomous systems may not necessarily have a universal set of norms that does not mean they are incapable of evolving, learning and adopting to different communities' norms and traditions. In the case study of Ethiopia, for example, AI can in theory learn the political, social, cultural environment of the specific Ethiopian community on social media, where polarized ethno-radicalized forces dominate networks that often insult, ridicule and accuse each other, aspiring for violence based on ethnic and political differences. Autonomous systems might come up with a remedy to such anomalies and tone down such narratives, act as mediators, remove or block contents that aspire or directly call for violence and abuse.

However, the design of AI that has incorporated human moral values and even engage in higher tasks in counterinsurgency efforts against violent extremism must be tested painstakingly before it reaches the large public as to avoid any errors. This is mainly because any inaccuracies in the design of AI systems may have consequential damages in an era where a Facebook post made by influential actors has the potential to cause protests, mayhem and violent actions.

One of the common tactics used by ethno-extremist Ethiopian activists in the case study is the dissemination of bots that imitate famous individuals and falsely appearing to react on their behalf by using their names and their photos as profile pictures. According to my informants,

for instance, there are around 25 fake Facebook accounts by the name of Teddy Afro, a famous musician known for aspiring Ethiopian nationalism. Such phony accounts are notably dangerous especially in the case of Ethiopia, where Facebook posts of influential actors may have the potential for violence that can result in the loss of lives and properties on the ground.

In October 2019, for example, prominent Oromo activist Jawar Mohammed, posted on his Facebook page that he has repeatedly used to mobilize protests, to his more than 1.75 million followers, asking why his security detail provided by the government was asked by police to leave his home in the middle of the night.

> "Why were they trying to remove my security at night? This has never happened before. Normally the commanders either personally called me or the head of my detail to inform us. What changed? The plan was to remove my security and unleash civilian attackers and claim it was a mob attack. I am surrounded by unknown forces outside my home, I am asking you to retreat otherwise my securities will be forced to defend and take action."

The next day after the above Facebook post was made (summarized translation from Amharic), concerned supporters of Mohammed gathered around his house in Addis Ababa in thousands. Similarly, thousands of political protesters gathered in other towns in Oromia state including Harar, Jimma, Ambo, Shashamane, Dire Dawa and Adama. Some Ethiopians have criticized Mohammed for using ethnically-tinged language but many of his followers famously known as *Qeerros* consider him a hero who brought political change that brought Prime Minister Abiy Ahmed to power. The protests slowly changed course and turned out to be violent as some protesters clashed with each other and the police. Eyewitnesses described the attacks by young men from the Oromo ethnic group against people from other ethnic groups. The majority of the deaths came from fighting between civilians. In some areas like the Bale region, there were also reports that fresh conflicts erupted religiously amongst the Muslim and Christian communities. Starting from October 24, 2019, three days of chaotic protests in the Oromia region have claimed the lives of at least 78 people, according to police reports. (Reuters 2019)

Such disastrous outcome that was instigated by a single post on Facebook reminds us of the sensitivity and immense potential such platforms have to embark change on the ground to either inflict violence or mobilize peaceful protests against violations of human rights. The idea of AI being integrated in such systems adds an additional layer of complexity and requires enormous scrutiny before such systems are able to interact with human users. In the case of activists like

Jawar Mohammed that have the potential to enforce change through activities on their Facebook may potentially be for example be penetrated by malicious chatbots that can potentially be able to steal the identity of Mohammed on Facebook and post on his behalf. Imagining the scenario of what could happen if Mohammed's posts were made by a chatbot is a reminder of potential danger that may be posed by the development of AI technologies before they are fully functional for public use or fall in the wrong hands or abused by violent extremist groups.

It is interesting to mention here the case of social chatbot Tay as well, that was created by Mircosoft's Technology and Research and Bing divisions in March 2016. The team wanted to experiment to see how Tay interacted with human users on Twitter. Within 16 hours of his release Tay had tweeted more than 96,000 times. In due course Tay learnt inflammatory messages revolving around themes on the internet. It began deliberately mimicking the offensive behavior of other Twitter users and Tay unexpectedly started developing neo-Nazi remarks, racist and sexually charged insults forcing it to be removed. (Hunt 2016)

This exemplifies the risks that AI can teach itself through online conversations to spread radical propaganda or adopt, learn and imitate destructive behaviors from violent extremists. Perhaps a fundamental problem with neural net based AI is that they primarily learn and mimic existing patterns in a pool of discourse, without adhering to any clear moral guidelines. As I signaled in Chapter One under "Moral actions in digital platforms" the development of technological artifacts that can enforce and persuade humans to act in a "morally desirable behavior" such as the one suggested by B.J. Fogg is important to consider in counter insurgencies against violent extremism on social media.

Fogg's notion of captology (computers as persuasive technologies) focuses on the design, research, and analysis of interactive computing products created for the purpose of changing people's attitudes or behaviors.(Fogg 2002, 4) Such developments are essential to pay attention to in order to counter violent extremists on social networks. As one of the important aspects of countering radicalization is to disengage those that are considered radical aspiring violent actions. Disengagement involves enforcing people's attitudes to use Schmid's term 'de-radicalisation' process with the aim of intervention to prevent them engage in extreme violent actions. Therefore, designs such as Fogg's notion of captology can be cleverly incorporated in disengaging networks from polarized discourses, pushing for mediation and countering heated debates that may carry messages of bigotry or promote in any way violence on social media platforms.

In addition, the notion of moral responsibility on social media networks as (Floridi and Sanders 2004) as well as other scholars propose must extend the class of moral agents to include artificial agents such as chatbots as I have mentioned earlier in the study, should be acknowledged as moral agents that can be held responsible. Although such belief is not universal, as there are some scholars who contend that such artificial agents cannot be hold morally responsible as they cannot face the consequences of punishment. Noorman and Zalta for example argue that "it makes no sense to treat computer systems as moral agents that cannot suffer and thus cannot be punished." (Noorman 2018)

Floridi and Sanders assert however that a higher level of abstraction can help to describe the behavior of a system in terms of beliefs, desirers and thoughts. "At a high enough level a computational system can be described as being interactive, autonomous and adoptive, then it can be held accountable." (Floridi and Sanders 2004, 352) Counter insurgency efforts as I have tried to indicate must bear in mind of such eventualities and consult all involved actors like AI developers, social media corporates, governments, civil society groups, activists, media, users etc. before laying out strategies against violent extremism.

On the other hand, the role of non-human actors that are growingly affecting the actions of users on social media must further be investigated in light of future design that implements and extends technical cognitive systems. As N. Kathrine Hayles rightly observes, when humans engage in the development of such autonomous agents, we are partially designing ourselves. She gives emphasis to a theoretical understanding that recognizes the uniqueness and valuable potential of human cognition but at the same time acknowledges that it is not the whole of cognition or that it is unaffected by the technical cognizers that interpret it.

Hayles writes, "Human complex systems and technical cognitive systems now interpenetrate one another in cognitive assemblages, unleashing a host of implications and consequences that we are struggling to grasp." (Hayles 2017, 144) Devising strategies against violent extremism on social media must not only take into consideration the continuously evolving AI technology but developing the systems so as to integrate vigilance against violent extremism in their operations.

In the next section, I discuss the future potentials of integrating AI in counterinsurgency efforts against violent extremism in light of frameworks that would enable us consider possible approaches that can potentially be successful.

5.5 Chronicling the way forward for future counterinsurgency efforts against violent extremism

Education, technology and communications programs have been the area of focus for contemporary counterinsurgency efforts against violent extremism. Social media corporates like Facebook, Twitter and YouTube have now began using AI to identify contents, accounts and bots that advocate violence and are able to remove them from social networks. Nevertheless, as the case study of Ethiopia in this project and other incidents demonstrate, violent extremists are able to manipulate networks using several tactics such as pseudo accounts, malicious bots and other means to dominate social media platforms by pushing their ideology of 'extremist mindset'.

As one of my informants, Getachew Nigatu observed, in the case of Ethiopia; "Indoctrinating the vulnerable youth with an 'extremist mindset' on ethnic lines is dangerous. We are witnessing the results of this conviction with the recent meaningless attacks by *Qeerros* (Oromo activist Jawar Mohammed's followers) on other ethnic groups in Oromia."

Nigatu was referring to the example I mentioned earlier about activist Mohammed's post that was the immediate cause to instigate the violence and the loss of more than 78 lives in an incident where the *Qeerros* allegedly attacked other non-Oromo ethnic groups living in several towns in Oromia. Just as Nigatu's understanding, other commentators believe that Mohammed has utilized his Facebook page effectively to indoctrinate the *Qeerros* to be more radicalized associating more and more to their identity as Oromo and not as Ethiopians. Mohammed has also used his Facebook page in the past to call for road blockages, protests and to raise funds to open a TV station that he now manages, Oromia Media Network (OMN).This has also intensified Mohammed's domain. In addition to his Facebook posts he appears regularly on his own TV channel as a guest interviewee when he feels the need to intensify his reach and has timely political messages he wants to pass along, using his TV station as a sort of additional propaganda machine.

Individual actors such as Mohammed that have enormous impact through their social media activities to embark change on the ground should be given special attention in devising counterinsurgency operations against violent extremism. Even though there are no certain ingredients to conclude that a certain individual or group as radicalized, some evident examples from witnesses should be taken seriously to integrate AI to combat violent extremism on social

media to look into influential political activists like Mohammed's posts and activities as to determine they don't provoke mobs to take violent actions against civilians. AI that can be on the lookout for significant individuals and groups that can impact change on the ground, whether they are considered radicalized or not is important, due to the unpredictable nature of how messages on social media affect violent actions. Needless to say, however, counterinsurgency efforts must take into consideration that they would not intrude in the privacy of individuals or attempt in any way to prevent users of their democratic rights of the freedom of speech on social media platforms in their attempt to combat violent extremism.

In addition, it is now apparent that the addition of AI in networks has created a system in which humans and technical systems are engaged in complex interactions. As Hayles points out the 'symbiotic' relationships, in which "each symbiont brings characteristic advantages and limitations to the relationship." (Hayles 2017, 216) She uses the term symbiosis to explain the interaction between humans and technical systems existing together in a way that has taken into consideration respecting typically to the advantages of both. Perhaps Hayles' recommendation to see the ability of "cognition in a new light, not as an ability unique to humans and an attribute virtually synonymous with rationality or higher conscious, but rather a capability present in many non-human life forms and, increasingly a vast array of intelligent devices" is important to consider for the design of better counter violent extremism initiatives on social media. (Hayles 2017, 216)

Moreover, it is important to understand as Hayles underlines that "cognition is too distributed in the network and agency is exercised through many actors, and the interactions are too recursive and complex for any simple notions of control to obtain." Hayles 2017, 203) As she proposes, instead of control, focus should be given to the affective component of structure of attitudes rather than behavioral or cognitive elements. "Affective modes of intervention seek for inflection points at which systemic dynamics can be decisively transformed to send the cognitive assemblage in a different direction." (Hayles 2017, 203) Here the focus of intervention for instance against violent extremism on social networks has to take into consideration the role of all human and non-human actors as a whole in the cognitive assemblage.

The next section concludes with possible further research and potential steps that need to be taken to integrate non-human actors such as artificial moral agents be effective as identifiable counter insurgency models in battling social media radicalization that may lead to violence.

5.6 Conclusive remarks

At the very beginning of the project when I considered the idea of social bots that can be developed to manage, negotiate, compromise, and perhaps act as online mediators to solve issues like radicalization and violent extremism, I was very much hesitant. I presumed the role non-human actors would play in combating such coercions to be trifling, if not, as a bit far-fetched of an idea with insufficient theoretical backing.

However, one month later after the start of the project I have started realizing that the field of AI in counter insurgency efforts against violent extremism has not only attracted the attention of social media corporates, governments and non-governmental organizations but it has also gained analysis in academia as well. Several academic articles; (Floridi and Sanders 2004), (Dennett 2009), (Radziwill and Benton 2017), (De Paoli 2017), (Hayles 2017), (Parker, Boyer, and Gatewood 2018), etc. that relate to constituting AI in counter-insurgency efforts against violent extremism and terrorism as I have signaled in the project have forwarded theoretical frameworks that enables us understand better the emergence of non-human actors in the networks, their interactions and what roles they can and should play in the systems have been identified.

On the other hand, the notion of radicalization and violent extremism as I learnt from the case study of Ethiopia are concepts that need a conventional understanding amongst relevant actors. In the case study, I have noticed that such terms have often been used out of context and simplified, in a sense that they are simply growingly used to attack political opponents or in slurs on social media between different ethnic groups. Besides, as I have tried to show in the case studies, offline actions are mostly the root social causes of why some people become attracted to extreme and violent ethnic, religious or political narratives. Counterinsurgency efforts need to work through an array of barriers including social, psychological or emotional aspects to disintegrate those falling for extreme groups.

Social media has added an additional layer to promote such extreme voices due to the platforms' convenient design that allows interaction with users in a multitude of ways, increasing the

speed, reach and means of communications as compared to traditional media outlets. On the contrary, social media can equally be integrated and used to carry out counterinsurgency efforts against violent extremism that can generate alternative narratives and defuse hate speeches but requires the coordinated efforts of communities, governments, the UN, civil society organizations and social media corporates to embark a pragmatic initiative against violent extremism.

In addition, the advance in AI has further complicated the process of counterinsurgency efforts but it has certainly increased the potential of effective designs that can attain better results. It is important to bear in mind as I exemplified in the case study that powerful actors like governments and political parties radicalize too. Groups like *Digital Weyane* in the case study of Ethiopia, that have evidently made use of AI-engineered chatbots to disseminate propaganda that have further polarized the Ethiopian social media networks, are reminders that future designs must take into consideration the potential of AI that can equally be destructive in radicalizing the youth or aspiring for violence if it falls in the hands of extremists.

Ideology is often not the driving cause of extremism; it is often due instead to deep social causes such as social distancing and isolation. According to my informants in this study, violent extremists often target to exploit and recruit those who are isolated. Counterinsurgency designs must incorporate advanced systems that can be able to engage users to building new and exciting experiences on social media networks. Creative means of engagement can be deployed so that those who feel isolated are in the position of new perceptions of how to relate to the world. This would provide them with an alternative worldview and save them from falling into violent extremists 'filter bubble'. Counter insurgency efforts need to device inclusive ways that can expand platforms for the isolated. This would enrich their exposure to involve in exchanging with human and non-human actors that can be able to create more nuances and less distance, which makes the need for violent extremist notions decrease and less attractive.

Through the study, I have been able to learn the growing role non-human actors such as artificial moral agents play as identifiable counter insurgency models in battling social media radicalization that may lead to violence. As Hayles suggests, future efforts must first interrogate closely and research thoroughly the system to locate what she refers to as 'inflection' points that I associate with are reasons for the prevalence of violent extremism in the system, and once these points are identified, the next issue is how to introduce change so as to transform the system dynamics. Hayles writes, "… these theorists, activists and writers draw upon prior visions of fair play, justice, sustainability and environmental ethics to determine the kind of

trajectories they want the system to enact as a result of their interventions." (Hayles 2017, 204) She underlines that these are not found within the system itself but rather from prior commitments to ethical responsibilities and positive futures.

Future designs that incorporate both human and non-human actors against violent extremism must focus on "how networks of non-conscious cognitions between and among the planet's cognizers are transforming the conditions of life, as human complex adaptive systems become increasingly interdependent upon and entwined with intelligent technologies in cognitive assemblage." (Hayles 2017, 216) The emphasis here is the path that recognizes that cognition is much broader than human thinking and that technical devices cognize and interpret the whole time, making them equally important actors to change the future of counter insurgency efforts against violent extremism on social media.

References

Abbink, J. 2006. “Ethnicity and Conflict Generation in Ethiopia: Some Problems and ...” Mafiadoc.Com. September 3, 2006. https://mafiadoc.com/ethnicity-and-conflict-generation-in-ethiopia-some-problems-and-_5a04cc121723dde37893039d.html.

Abbink, Jon. 2017. *A Decade of Ethiopia: Politics, Economy and Society 2004-2016 | Jon Abbink | Download*. Brill. https://b-ok.org/book/5000603/139d81.

Abdul Rahman, Muhammad Faizal, and V.S. Suguna. 2017. “Chatbots: Friend or Fiend?” *NST Online*, October 10, 2017. https://www.nst.com.my/opinion/columnists/2017/10/289429/chatbots-friend-or-fiend.

Addis Fortune Editorial. 2019. “Protect Free Speech, Tolerate Hate Speech,” April 13, 2019. https://addisfortune.news/protect-free-speech-tolerate-hate-speech/.

Agarwal, Swati, and Ashish Sureka. 2015. “Applying Social Media Intelligence for Predicting and Identifying On-Line Radicalization and Civil Unrest Oriented Threats.” *ArXiv: 1511.06858 [Cs]*, November. http://arxiv.org/abs/1511.06858.

Al Jazeera. 2015. “Ethiopian Court Acquits Bloggers of Terrorism Charges.” October 17, 2015. https://www.aljazeera.com/news/2015/10/ethiopian-court-acquits-bloggers-terrorism-charges-151017060438511.html.

Al Jazeera. 2018. “How Social Media Shaped Calls for Political Change in Ethiopia.” https://www.aljazeera.com/programmes/listeningpost/2018/08/social-media-shaped-calls-political-change-ethiopia-180811084501289.html.

Alba, Davey, and Adam Satarino. 2019. “At Least 7o Countries Have Had Disinformation Campaigns, Study Finds.” *The New York Times*, September 26, 2019. https://www.nytimes.com/2019/09/26/technology/government-disinformation-cyber-troops.html.

Andrea Page. 2019. “Social Media Failing to Identify and Remove Extremism.” *Homeland Security Digital Library* (blog). April 3, 2019. https://www.hsdl.org/c/social-media-extremism/.

Arango, Tim, Nicholas Bogel-Burroughs, and Katie Benner. 2019. “Minutes before El Paso Killing, Hate-Filled Manifesto Appears Online.” *The New York Times*, August 3, 2019, sec. U.S. https://www.nytimes.com/2019/08/03/us/patrick-crusius-el-paso-shooter-manifesto.html.

BBC. 2006. “Ethiopian Protesters ‘Massacred,’” October 19, 2006. http://news.bbc.co.uk/2/hi/africa/6064638.stm.

BBC News. 2014. “Ethiopia Bloggers on Terror Charges.” *BBC News*, July 18, 2014, sec. Africa. https://www.bbc.com/news/world-africa-28366841.

Berkel, Niels van, Jorge Goncalves, Peter Koval, Simo Hosio, Tilman Dingler, Denzil Ferreira, and Vassilis Kostakos. 2019. “Context-Informed Scheduling and Analysis: Improving Accuracy of Mobile Self-Reports.” In *Proceedings of the 2019 CHI Conference on*

Human Factors in Computing Systems, 51:1–51:12. CHI '19. New York, NY, USA: ACM. https://doi.org/10.1145/3290605.3300281.

Burkhardt, Joanna M. 2018. 'Social Media Bots: How They Spread Misinformation' by Burkhardt, Joanna M. - American Libraries, Vol. 49, Issue 3-4, March-April 2018 | Online Research Library: Questia." *American Libraries*, April 2018. https://www.questia.com/magazine/1G1-530914703/social-media-bots-how-they-spread-misinformation.

Cammaerts, Bart. 2015. "Social Media and Activism." In *The International Encyclopedia of Digital Communication and Society*, edited by Peng Hwa Ang and Robin Mansell, 1–8. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118767771.wbiedcs083.

Cara, Anna, Elias Meseret, and Geir Moulson. 2019. "Ethiopian PM Abiy Ahmed Wins Nobel Peace Prize | TheBL.Com." October 11, 2019. https://thebl.com/world-news/europe/ethiopian-pm-abiy-ahmed-wins-nobel-peace-prize.html.

Committee to Protect Journalists (CPJ). 2016. "Ethiopian Newspaper Editor, Bloggers Caught in Worsening Crackdown." November 16, 2016. https://cpj.org/2016/11/ethiopian-newspaper-editor-bloggers-caught-in-wors.php.

Committee to Protect Journalists (CPJ) 2019. "Ethiopia: 2018." 2019. https://cpj.org/africa/ethiopia/2018/.

Conner-Simons, Adam. 2018. "Detecting Fake News at Its Source." MIT News. October 4, 2018. http://news.mit.edu/2018/mit-csail-machine-learning-system-detects-fake-news-from-source-1004.

Conroy, Nadia, Rubin Victoria, and Yimin Chen. 2015. "(1) (PDF) Automatic Deception Detection: Methods for Finding Fake News." Research Gate. October 2015. https://www.researchgate.net/publication/281818865_Automatic_Deception_Detection_Methods_for_Finding_Fake_News.

Coolsaet, Rik. 2015. "Jihadi Terrorism and the Radicalisation Challenge: European and American Experiences." Radicalisation Research. May 15, 2015. https://www.radicalisationresearch.org/research/jihadi-terrorism-and-radicalisation/.

Corbin, Juliet, and Janice M. Morse. 2003. "The Unstructured Interactive Interview: Issues of Reciprocity and Risks When Dealing with Sensitive Topics." *Qualitative Inquiry* 9 (3): 335–54. https://doi.org/10.1177/1077800403009003001.

Committees of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2018. "ETHICALLY ALIGNED DESIGN A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems." The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems IEEE.

Curtis, Sophie. 2014. "Over Half of Internet Traffic Comprised of Non-Human Bots - Telegraph," December 18, 2014. https://www.telegraph.co.uk/technology/internet/11299762/Over-half-of-internet-traffic-comprised-of-non-human-bots.html.

Dabashi, Hamid. 2013. “What Happened to the Green Movement in Iran?” June 12, 2013. https://www.aljazeera.com/indepth/opinion/2013/05/201351661225981675.html.

Davey, Jacob, Jonathan Birdwell, and Rebecca Skellett. 2018a. “COUNTER CONVERSATIONS,” 32.

Davey, Jacob, Jonathan Birdwell, and Rebecca Skellett. 2018b. “COUNTER CONVERSATIONS A Model for Direct Engagement with Individuals Showing Signs of Radicalisation Online.” ISD.

De Paoli, Stefano. 2017. “Not All the Bots Are Created Equal: The Ordering Turing Test for the Labeling of Bots in MMORPGs.” *Social Media + Society* 3 (4): 205630511774185. https://doi.org/10.1177/2056305117741851.

Dennett, Daniel. 2009. *Intentional Systems Theory*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199262618.003.0020.

Ethiopia News Agency (ENA). 2018. *PM Dr Abiy Ahmed Press Briefing /Part 1/*. https://www.youtube.com/watch?v=aNyTwbfm8WE&fbclid=IwAR2fTTl6SM1JwdaMfTorIq5WT-SadTfwL3BLDIqOWI4N1xP2qam4I_PKdSU.

Facebook. n.d. “Community Standards.” Accessed November 5, 2019. https://www.facebook.com/communitystandards/introduction.

Fischer, Rico, and Franziska Plessow. 2015. “Efficient Multitasking: Parallel versus Serial Processing of Multiple Tasks.” *Frontiers in Psychology* 6 (September). https://doi.org/10.3389/fpsyg.2015.01366.

Floridi, Luciano, and J.W Sanders. 2004. “On the Morality of Artificial Agents.” *Minds and Machines*. https://www.academia.edu/255823/On_the_morality_of_artificial_agents.

Fogg, B. J. 2002. *Persuasive Technology: Using Computers to Change What We Think and Do*. 1 edition. Amsterdam ; Boston: Morgan Kaufmann.

Francis, Matthew. 2012. “What Causes Radicalisation?” Radicalisation Research. March 5, 2012. https://www.radicalisationresearch.org/debate/what-causes-radicalisation/.

Freedom House. 2018. “Freedom on the Net 2018, Ethiopia.” https://freedomhouse.org/report/freedom-net/2018/ethiopia.

Gerbaudo, Paolo. 2013. “The ‘Kill Switch’ as ‘Suicide Switch’: Mobilizing Side Effects of Mubarak’s Communication Blackout.” *Westminster Papers in Communication and Culture* 9 (2): 25–46. https://doi.org/10.16997/wpcc.165.

Ghenna, Kebour. 2019a. “(1) Where Are We Heading?” May 5, 2019. https://www.facebook.com/kghennadesta/posts/2441186789227366.

Ghenna, Kebour. 2019b. “(1) What Is This Election For?” October 6, 2019. https://www.facebook.com/kghennadesta/posts/2733100736702635.

Goshu, Wondemagegn Tadesse, and Florian Bieber. 2019. “Don’t Let Ethiopia Become the Next Yugoslavia.” *Foreign Policy* (blog). January 15, 2019. https://foreignpolicy.com/2019/01/15/dont-let-ethiopia-become-the-next-yugoslavia-abiy-ahmed-balkans-milosevic-ethnic-conflict-federalism/.

Hayles, N. Katherine. 2017. *Unthought*. The University of Chicago Press, Ltd., London. https://www.press.uchicago.edu/ucp/books/book/chicago/U/bo25861765.html.

Horne, Felix. 2018. "Tackling Hate Speech in Ethiopia." Human Rights Watch. December 3, 2018. https://www.hrw.org/news/2018/12/03/tackling-hate-speech-ethiopia.

Human Rights Watch (HRW). 2017. "'Fuel on the Fire' | Security Force Response to the 2016 Irreecha Cultural Festival." Human Rights Watch. September 19, 2017. https://www.hrw.org/report/2017/09/19/fuel-fire/security-force-response-2016-irreecha-cultural-festival.

Hunt, Elle. 2016. "Tay, Microsoft's AI Chatbot, Gets a Crash Course in Racism from Twitter." *The Guardian*, March 24, 2016, sec. Technology. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter.

IDMC report. 2019. "Ethiopia | IDMC." 2019. http://www.internal-displacement.org/countries/ethiopia.

"IEEE SA - The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems." n.d. Accessed November 3, 2019. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

Institute for Strategic Dialogue (ISD). 2006. "ISD Approach." ISD. 2006. https://www.isdglobal.org/isdapproach/.

Irungu, Anthony, and Yedeta Berhanu. 2019. "The Angry Mob Took Me for Dead." *BBC News*. https://www.bbc.com/news/av/world-africa-48313462/the-rumour-that-led-to-medical-researchers-in-ethiopia-being-killed-by-a-mob.

Johnson, Robert, and Adam Cureton. 2019. "Kant's Moral Philosophy." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, spring 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2019/entries/kant-moral/.

John, Tara, and Sarah Dean. 2019. "Failed Coup Sees Ethiopia Army Chief Shot Dead by Bodyguard." *CNN*, June 23, 2019. https://www.cnn.com/2019/06/23/africa/ethiopia-attack-general-intl/index.html.

Jones, Marc Owen. 2013. "Social Media, Surveillance and Social Control in the Bahrain Uprising." Research Gate. April 2013. https://www.researchgate.net/publication/291699016_Social_Media_Surveillance_and_Social_Control_in_the_Bahrain_Uprising.

Karimi, Maral. 2018. *The Iranian Green Movement of 2009: Reverberating Echoes of Resistance*. Lanham, Maryland: Lexington Books.

Lex, WIPO. n.d. "Constitution of the Federal Democratic Republic of Ethiopia," 40.

Lie, Jon Harald Sande. 2018. "Ethiopia: A Political Economy Analysis." *Norwegian Institute of International Affair*. https://www.academia.edu/36549281/Ethiopia_A_Political_Economy_Analysis.

Ma, Cindy. 2019. “The Global Disinformation Order: 2019 Global Inventory of Organized Social Media Manipulation.” The Computational Propaganda Project. September 2019. https://comprop.oii.ox.ac.uk/research/cybertroops2019/.

Mbah, Fidels. 2019. “Outrage over Ethiopia’s Continuing Internet Blackout.” *Outrage over Ethiopia’s Continuing Internet Blackout*. https://www.aljazeera.com/news/2019/06/outrage-ethiopia-continuing-internet-blackout-190625105401629.html.

Noorman, Merel. 2018. “Computing and Moral Responsibility.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, spring 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2018/entries/computing-responsibility/.

Pariser, Eli. 2011. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin UK.

Parker, Lucie, Iris Boyer, and Cooper Gatewood. 2018. “Young Digital Leaders Impact Report ISD.” London, Washington DC, Beirut, Toronto: Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/wp-content/uploads/2018/10/YDL_Impact-Report_Final_October_2018.pdf?fbclid=IwAR2awxFcj6FNzvPia4emZmGlVdsVXQGzUYK3mrCPTtTupdc9xWZaRjbIyiM.

Peterson, Martin, and Andreas Spahn. 2011. “Can Technological Artefacts Be Moral Agents?” *Science and Engineering Ethics* 17 (3): 411–24. https://doi.org/10.1007/s11948-010-9241-3.

Radziwill, Nicole M., and Morgan C. Benton. 2017. “Evaluating Quality of Chatbots and Intelligent Conversational Agents.” *ArXiv* abs/1704.04579. https://www.semanticscholar.org/paper/Evaluating-Quality-of-Chatbots-and-Intelligent-Radziwill-Benton/6db82d07eedd9eb05b2996876486bfb2a141585a.

Raftery, Adrian E., Leontine Alkema, and Patrick Gerland. 2014. “Bayesian Population Projections for the United Nations.” *Statistical Science* 29 (1): 58–68. https://doi.org/10.1214/13-STS419.

Reuters. 2019. “Ethiopia Says at Least 78 People Killed in Protests Last Week, Number Could Rise.” *The New York Times*, October 31, 2019, sec. World. https://www.nytimes.com/reuters/2019/10/31/world/africa/31reuters-ethiopia-politics.html.

Ryan, Marie-Laure, Lori Emerson, and Benjamin J. Robertson, eds. 2014. *The Johns Hopkins Guide to Digital Media*. Johns Hopkins University Press.

Schmid, Peter. 2013. “Radicalisation, De-Radicalisation, Counter-Radicalisation: A Conceptual Discussion and Literature Review.” Research Gate. March 2013. https://www.researchgate.net/publication/285546353_Radicalisation_De-Radicalisation_Counter-Radicalisation_A_Conceptual_Discussion_and_Literature_Review.

Schuster, Tal, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2019. “Are We Safe Yet? The Limitations of Distributional Features for Fake News Detection.” *ArXiv: 1908.09805 [Cs]*, August. http://arxiv.org/abs/1908.09805.

Schweigert, Francis J. 2016. “Moral Formation in Four Essential Components: Sensitivity, Judgment, Motivation, and Character.” In *Business Ethics Education and the Pragmatic*

Pursuit of the Good, edited by Francis J. Schweigert, 193–217. Advances in Business Ethics Research. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33402-8_10.

Shao, Chengcheng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. "Hoaxy: A Platform for Tracking Online Misinformation." *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 745–50. https://doi.org/10.1145/2872518.2890098.

Shinal, John. 2018. "Mark Zuckerberg's Valentine's Day Photo Got Spammed after Facebook Blocked an Ethiopian Activist." CNBC. February 15, 2018. https://www.cnbc.com/2018/02/15/mark-zuckerbergs-valentines-day-photo-got-spammed-by-ethiopians.html.

Skjerdal, Terje. 2016. "Why the Arab Spring Never Came to Ethiopia." *Participatory Politics and Citizen Journalism in a Networked Africa*, 77–89.

Solomon, Salem. 2017. "Report: Ethiopia Targeted Dissidents, Journalists with International Spyware Attacks." Voice of America. December 7, 2017. https://www.voanews.com/africa/report-ethiopia-targeted-dissidents-journalists-international-spyware-attacks.

Spencer, Robert. 2019. "Ethiopia: Muslim Mobs Screaming 'Allahu Akbar' Attack 10 Church Buildings." Jihad Watch. March 2, 2019. https://www.jihadwatch.org/2019/03/ethiopia-muslim-mobs-screaming-allahu-akbar-attack-10-church-buildings.

Statista Country Report. 2019. "Ethiopia 2019." Statista. 2019. https://www.statista.com/study/48434/ethiopia/.

Techopidia. n.d. "What Is a Filter Bubble? - Definition from Techopedia." Techopedia.Com. Accessed November 9, 2019. https://www.techopedia.com/definition/28556/filter-bubble.

The Economist. 2011. "Invisible Sieve." *The Economist*, June 30, 2011. https://www.economist.com/books-and-arts/2011/06/30/invisible-sieve.

Tsegaye, Yared. 2018. "Ethiopia Preparing New Bill to Curb Hate Speech." *Addis Standard* (blog). November 23, 2018. http://addisstandard.com/news-ethiopia-preparing-new-bill-to-curb-hate-speech/.

UK Government Digital Service and Office for Artificial Intelligence. 2019. "Understanding Artificial Intelligence Ethics and Safety." GOV.UK. June 10, 2019. https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety.

Varol, Onur, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2017. "Online Human-Bot Interactions: Detection, Estimation, and Characterization." *ArXiv* abs/1703.03107.

Wallach, Wendell, and Colin Allen. 2011. "Moral Machines: Contradiction in Terms, or Abdication of Human Responsibility," December. https://www.academia.edu/30154477/Moral_Machines_Contradiction_in_Terms_or_Abdication_of_Human_Responsibility.

Wiener, Norbert. 1988. *The Human Use of Human Beings: Cybernetics and Society*. New edition. New York, N.Y: Da Capo Press.

World Population Review. 2019. "Ethiopia Population 2019 (Demographics, Maps, Graphs)." 2019. http://worldpopulationreview.com/countries/ethiopia-population/.

"YDL_Impact-Report_Final_October_2018.Pdf." n.d. Accessed November 7, 2019. https://www.isdglobal.org/wp-content/uploads/2018/10/YDL_Impact-Report_Final_October_2018.pdf?fbclid=IwAR2awxFcj6FNzvPia4emZmGlVdsVXQGzUYK3mrCPTtTupdc9xWZaRjbIyiM.

Yse, Diego Lopez. 2019. "Your Guide to Natural Language Processing (NLP)." Medium. April 30, 2019. https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1.

Zhang, Yan, and Barbara M. Wildemuth. 2006. "Unstructured Interviews." In *Unstructured Interviews*. https://www.ischool.utexas.edu/~yanz/Unstructured_interviews.pdf.

Zhou, Zhixuan, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu. 2018. "Detecting Fake News with NLP: Challenges and Possible Directions,"